

Einstein is dead

Until its next revolution, much of the glory of physics will be in engineering. It is a shame that the physicists who do so much of it keep so quiet about it.

Once upon a time there was (and still is) a multinational manufacturer of sheet metal whose researchers realized they could improve the reliability of its production processes. By solving the equations of heat transfer for the company's rolling presses, and testing the solutions on scale models, they significantly reduced the margin of error in the thickness of the rolled sheet. Their prime customer, a manufacturer of metal cans, was delighted. Millions of pounds were saved in materials and rejected products, and (maybe) the reduced costs were passed on to the customer. Maybe, too, the physicists got bonuses.

How very far removed from the special theory of relativity and the world of quantum mechanics — the parallel revolutionary paradigms on which most of twentieth-century physics and related technologies were based. Now, 100 years after Einstein's first pioneering papers in those disciplines, physicists worldwide are rightly going to town, with conferences, artistic commissions and games. (For some of *Nature's* own partying, see pages 200 and 213.) They are doing their utmost to celebrate in the face of the relentless promotion of biology as the exciting science of the current century and despite declining interest in physics amongst the young.

Einstein is not only the patron saint of physics but also an icon of integrity and scientific pursuit for its own sake — and, for the wider public, an appealing elderly gentleman. Small wonder, then, that UK and Irish physicists opted to call 2005 'Einstein year', rather than the 'Year of physics'. But Einstein is long gone. His ideas, his style and his legacy still inspire, but his rejection of the quantum picture of reality and his dreams of the unification of forces have been replaced by the acceptance and exploration of quantum entanglement and highly esoteric (albeit potentially profound) attempts to derive twentieth-century laws from a deeper paradigm for the structure of space-time.

To hang a 'Year of physics' so centrally on Einstein is to miss the key lessons of the metal manufacturer: that physics is not only central to our understanding of the Universe (just what are dark matter and dark energy?), but is also central to making useful and sometimes

inspiring things. Sheet metal is at the more prosaic end of the spectrum. At the other end, Steve Jobs, head of Apple, said at last week's launch of the latest iPod: "Most people make the mistake of thinking design is what it looks like ... Design is how it works." In other words, sexy design is also about sexy engineering and the sexy science behind it.

And listen to theoretical physicist Michael Berry of the University of Bristol, UK, launching the competition "Physics for taxi drivers" (*Physics World* December 2004, p. 15; <http://physicsweb.org/articles/world/17/12/2>). He recalls how a description of a CD player and a satellite navigation receiver convinced a cab driver that physics is interesting. The worry is not so much that people cannot understand the relevance of physics — and credit to the 'World Year of Physics 2005' organizers for a poster competition for 10–16-year-olds to celebrate that. The worry is that in universities, and especially in schools, there is so little emphasis and imagination, either this year or ever, in celebrating physics' relevance and, more importantly, sending the right career signals to young people.

Many young people today are as capable as previous generations of being inspired by the challenge of making things: engineering with unbelievable precision in the face of quantum uncertainties, creating elegance in functional design, and delivering innovative and useful — or even socially transforming — everyday things. *Nature's* pages have included their share of the foundations of twenty-first-century manufacturing, with advances in the quantum control of atomic and molecular states, quantum information and optoelectronics (see page 292, for example).

Some of the authors of those papers have interesting engineering careers ahead of them. As surveys by learned societies repeatedly show, a large proportion of physics graduates find fulfilling and well paid employment in engineering and information technology. Those same societies, and governments and physicists generally, repeatedly fail to get that message across to the public or to kids in schools. Yet that is surely a more important challenge this year than reiterating in depth, appropriately but ineffectually, that Einstein was great. ■

Tales of the unexpected

Unfettered research sometimes leads to highly serendipitous discoveries.

X-rays, penicillin and the World Wide Web — these are often-cited examples of how unshackled science can lead curious researchers to unforeseeable discoveries. But such discoveries quickly get taken for granted, so it is good occasionally to highlight new examples of the unexpected.

One can be found in research into fitness and exercise. Few if any people could have anticipated the results that poured out of the long-term, US-based HERITAGE Family Study into exercise training. HERITAGE represents research at its most basic: engage large numbers of people, make them do fitness training for 20 weeks, gather diverse physiological, biochemical and genetic data with no hypothesis in mind, and see what you get.

The researchers found an astonishing variability in our individual

responses to training (see page 188). Some participants got much fitter, others hardly at all. And the study shook assumptions about the presumed risk factors for common diseases that are particularly prevalent among couch potatoes, such as cardiovascular disease and type 2 diabetes. In some outliers the overall profile of presumed risk factors was dramatically high, for others quite small.

Meanwhile, a UK-based study on athletes has led directly to clinical trials of a drug targeted to a particular 'fitness gene' in highly vulnerable groups of patients — premature babies, for example — whose very life depends on their efficient use of oxygen, just like an athlete's ability to win medals.

Scholars of science, policy-makers and those who simply enjoy the highways and byways of research, take note. ■

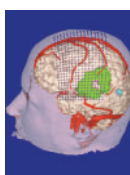




Life lessons
Court rules against
anti-evolution
stickers on textbooks
p182



Potential energy
Seismologists focus
on ocean fault's
unspent power
p183



Ethical headache
Brain scans pose
dilemmas for
medical researchers
p185



Taking a stand
Iranian scientist
joins race to become
president
p186

Titan team claims just deserts as probe hits moon of *crème brûlée*

Alison Abbott, Darmstadt

Scientists were ecstatic last weekend as Titan, Saturn's largest moon, dramatically revealed itself to have the atmosphere-bearing, hydrocarbon-based landscape they had anticipated.

But Huygens, the European Space Agency (ESA) spacecraft that successfully parachuted through Titan's atmosphere on 14 January, has also revealed its share of surprises. The high-risk mission had functioned better than its designers had dared to expect — and they quickly reported that the moon looked even more interesting than they had hoped.

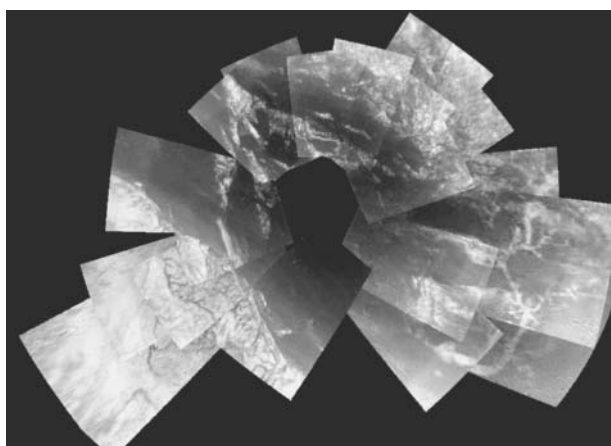
The probe successfully transmitted data from its six instrument packages during a 2.5-hour descent through Titan's atmosphere, and for more than 70 minutes after it landed.

Scientists waiting anxiously for the data to arrive at the European Space Operations Centre (ESOC) in Darmstadt, Germany, hugged each other when the first signals arrived during the morning, showing that the mission, 20 years in the planning and execution, was functioning. Many had worked most of the previous night preparing for the data's arrival.

Their delight grew as they began to look at what the data — relayed in compressed form from the mother ship Cassini in the late afternoon — seemed to be telling them.

"My worst nightmare had been that Titan would turn out to be a bland object, a simple icy surface with no structure or variety," says John Zarnecki, principal investigator on the Surface Science Package at the Open University in Milton Keynes, UK. "But it is fascinating geologically — it has weather, a variegated surface with lots of features — and it keeps alive the idea that it has lakes made of liquid organic material."

It was 15 hours from the start of the data dump to the press conference that brought the world the first images of Titan. And they were perhaps the most intense times in the professional lives of the six Huygens investigators



Caught on camera: early images of Titan's surface from Huygens.

and their teams from institutes in France, Germany, Britain and the United States.

Celebration and concentrated work ran in parallel. High-ranking guests, including ministers and space-agency officials, banqueted in ESOC's main building as scientists devoured data in the building next door.

Aware of the pressing need to tell the world a story at short notice, yet knowing that the whole, nuanced tale would require years of data analysis, the teams sprang into action. They did quick scans of their entire data sets to see which individual instruments were delivering information that seemed to make immediate sense. Then all team members dived into analyses of these particular data to provide a taste of the stories to come.

Public image

"If we hadn't had a large public waiting for a story, we probably would not have worked this way," says Zarnecki. "We would have taken a more structured, disciplined approach, where team members would work through data from different parts of each experiment." But he adds that scientists "have changed over the past few years, and appreciate a duty to inform taxpayers of how their money is being spent".

Within a few hours of the landing, Marty Tomasko of the University of Arizona in Tucson had used data from his camera to select three pictures taken 16 km, 8 km and

a few centimetres from Titan's surface. Looking at the close-up photograph, experts at first thought they were seeing an old picture of Mars, with its flat, fog-free, boulder-filled landscape. "If this really were Titan, then we would be in big trouble," commented one NASA scientist at the banquet, confident that the files had been confused.

But later, as more information trickled through from the control room, the small scale of the picture was made clear. The boulders were actually just a few centimetres across, and possibly made of very hard ice. And spectroscopic results showed Titan's clouds to be high — around 20 km up — making for a clear lower atmosphere.

The intensity of the work — and of the celebrations — continued through the night in the science rooms. Marcello Fulchignoni of the Paris-Meudon Observatory was hurriedly transforming his team's results on the physical and electrical properties of Titan's atmosphere into sound. "The transformation doesn't have any scientific value, but it fires the imagination of the general public," he says.

Around midnight, Fulchignoni dived out to pick up the bottles of champagne, specially labelled with the logo of his experiment, which he had laid down when Cassini-Huygens was launched in 1997. By 2:30 a.m. Zarnecki, whose team was sharing the same room, had opened the malt whisky that he had won on a sweepstake to guess the exact time of Huygen's landing. Sometime in those heady hours, Zarnecki coined a description of Titan's soft but crusted surface that looks likely to stick: *crème brûlée*.

By the time the world's press began pouring through the ESOC gates at 10 a.m., many of the scientists were shuffling out like zombies. But the Huygens investigators managed to suppress their exhaustion and conduct interviews well into the early afternoon.

What happened next? A nap, perhaps? "Well, we worked a bit more in the afternoon, and then I took my team out for dinner in the evening," says Zarnecki. ■

ESA

Georgia court bans biology textbook stickers

Jessica Ebert, Washington

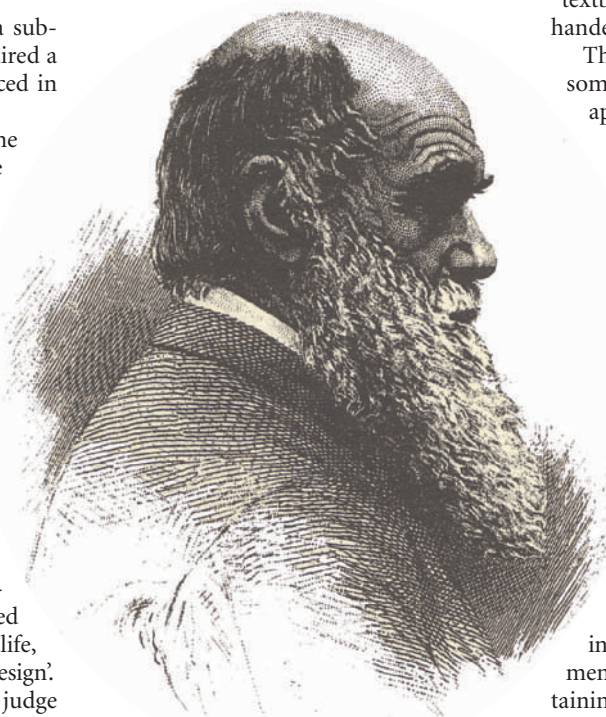
A court in Georgia has overturned a suburban Atlanta school policy that required a disclaimer about evolution to be placed in science textbooks.

Cobb County School District — the second largest in Georgia, with more than 100,000 students — started placing stickers in newly adopted high-school biology textbooks in the spring of 2002. The stickers read: “This textbook contains material on evolution. Evolution is a theory, not a fact, regarding the origin of living things. This material should be approached with an open mind, studied carefully and critically considered.”

But five local parents, backed by the American Civil Liberties Union of Georgia, sued the school district. They claimed the stickers inhibit the teaching of evolution and promote faith-based alternative views on the origins of life, including creationism and ‘intelligent design’.

On 13 January, Clarence Cooper, a judge at the district court in Atlanta, ruled that the sticker “misleads students regarding the significance and value of evolution in the scientific community”.

By placing stickers in science books, “the school board has effectively improperly entangled itself with religion by appearing to take a position”, he wrote in his ruling, and impressionable students “are likely to view the message on the sticker as a union of church and state”. This undermines the first amendment, the judge ruled. So, the stickers must go.



Sticky issue: creationists question Darwin's theory of evolution.

“This ruling is bigger than evolution,” says Jeffrey Selman, a computer programmer and one of the parents who filed the lawsuit. “This will defend keeping science education pure.”

The Cobb County Board of Education issued a written statement saying it was “disappointed” in the ruling, maintaining that

“textbook stickers are a reasonable and even-handed guide to science instruction”.

The board is expected to take legal advice sometime this week on whether or not to appeal, says Doug Goodwin, spokesman for Cobb County schools.

Textbook disclaimers are not a new tactic for spreading creationist views, says Eugenie Scott, head of the National Center for Science Education. The stickers have existed in one form or another since the 1970s and are popular with school boards across the country because they are cheap.

Although the Cobb County decision will not put an end to disclaimers entirely, there is a broader significance to the ruling, says Scott. “This is the first time that there has been a chance for a judge to rule on the softer forms of creationism.” She refers specifically to intelligent design, an intellectual movement that challenges evolution by maintaining that the complexity of the origin and diversity of life must have been created by an intelligent mind.

Scott sees intelligent design as an attempt to “repack creationism in a way that will avoid legal problems”. This will soon be tested in Dover, Pennsylvania — the first school district in the United States to add intelligent design to the science curriculum. Eleven parents have sued to overturn the policy and, according to Scott, the federal trial seems likely to begin this autumn. ■

All parties on edge as NIH delays open-access briefing

Erika Check, Washington

The US National Institutes of Health (NIH) has cancelled a briefing on the final version of its new policy for open access to scientific literature — leaving the plan’s supporters and opponents anxious about what happens next.

On 10 January, officials at the biomedical research agency alerted reporters and other interested parties that the NIH would unveil its open-access policy the next day. But that same evening, they abruptly cancelled the announcement, and declined to say when it will be rescheduled.

The plan, whose progress has been followed avidly by scientific publishers and many researchers, has already been outlined by NIH director, Elias Zerhouni. He has written that the agency will request, but not require, that NIH-funded researchers submit the final, peer-reviewed version of

their publications to the NIH. The agency would then make the manuscripts freely available after a specified time.

Zerhouni has previously said that the papers would be made public on the National Library of Medicine’s PubMed Central website no earlier than six months after the date of publication (E. A. Zerhouni *Science* 306, 1895; 2004). But multiple sources briefed on the new version of the plan last week say that the date has now changed to 12 months post-publication.

Sources close to Congress and the NIH speculated that the White House had scuttled the NIH announcement over concerns that the issue would complicate the confirmation hearings of Mike Leavitt, whom President George Bush has nominated as health secretary. Those hearings were set to be held on 18 and 19 January. But some questioned

this explanation, which wasn’t officially confirmed. This has left each side of the open-access debate worrying that the policy may now be revised in favour of their opponents.

“Obviously the policy could change — we’ve certainly heard that it may,” says Barbara Meredith, a vice-president at the Association of American Publishers, which opposes early, open release of all research findings.

“The fact that they’ve postponed the announcement gives us concern,” says Emily Sheketoff of the American Library Association, which supports quick, open access to literature. Sheketoff worries that her opponents may now influence the policy. The NIH “is not waiting to hear more from us”, she frets. “It is waiting to hear more from them.” ■



Seismic sensitivity: coral retains a record of changes in sea level, and hence of past earthquakes.

Indian Ocean fault line poses threat of further earthquakes

Emma Marris

Just weeks after the earthquake that triggered the devastating tsunami in the Indian Ocean, seismologists are turning their attention to nearby sections of the fault line that are also prone to rupture.

Researchers say that there could be further earthquakes both to the north and south of the event that occurred on 26 December, perhaps within decades, and that they might even be powerful enough to cause another tsunami.

The location of the epicentre of last month's earthquake is clear, but seismologists are uncertain where the northern edge of the rupture ended, says Kenji Satake, who studies tsunamis at the National Institute of Advanced Industrial Science and Technology in Tsukuba, Japan.

Some of the northern sections of the fault line may still be under enough strain to cause another earthquake of similar magnitude, says Satake, who hopes to go to the Andaman and Nicobar Islands in February to look for uplift of the land: no uplift would indicate that the fault has not yet ruptured in that area.

Likewise, some sections of the fault line south of the epicentre, which have definitely not yet ruptured, seem to be due for an earthquake.

Kerry Sieh, a geologist at the California Institute of Technology in Pasadena, has been investigating this area for years. He has a network of six instruments on a number of small islands off the coast of Sumatra, the exact locations of which are recorded by global positioning satellites. These reveal

that the islands are currently sinking by as much as a centimetre a year, as the continental plate on which they sit is slowly dragged down into the subduction zone by the oceanic plate. When these plates slip during an earthquake, the islands spring back up again as the pressure on the continental plate is released.

Sieh has been investigating historical records of earthquakes in the area by looking at the coral, which retains a record of the low-tide mark — coral can only live under water, so it dies off if the low-tide mark recedes, or grows if the low-tide mark rises. Sieh has used this to estimate changes in elevation to within a couple of centimetres over the past 50 to 100 years, and to look for larger rises associated with past earthquakes.

Sieh says that earthquakes in the region happen about once every 230 years. "We think we are getting close to the next one," he says. The last such event on the Mentawai islands was 145 to 170 years ago, for example. Energy from last month's earthquake will have placed extra strain on the fault, which is likely to hasten the next rupture.

As the Sumatra study progresses, Sieh and his team have been teaching the islanders about earthquake risks. "We have been trying to educate the villagers with posters and brochures, telling them how to build their structures," he says. "I cross my fingers that the people we started to educate took our advice and clambered up the hill when the tsunami hit."

Additional reporting by David Cyranoski.

For more on the tsunami visit:

► www.nature.com/news/infocus/tsunami.html

Pasteur board quits in bid to resolve crisis at troubled institute

Declan Butler, Paris

The board of directors of the Pasteur Institute in Paris has resigned en masse, in the latest twist in a bitter dispute between staff and management over the running of the institute.

The dispute centres on plans to relocate part of the prestigious research centre to a commercial zone on the city outskirts (see *Nature* 432, 788; 2004) — but it also involves what critics see as the brash management style of the institute's director-general, Philippe Kourilsky.

The institute's board has 16 elected members and four appointed by the government. The elected members are chosen every six years by its assembly — a sort of parliament of some 100 members.

Several board members have said they are unhappy with the direction that the biomedical research institute is taking. And the board has come under fire from the assembly and scientists for failing to intervene.

Last June, the assembly refused to approve the Pasteur's annual report prepared by the board, in protest at a looming funding crisis there (see *Nature* 430, 283; 2004), and in December hundreds of scientists picketed a board meeting, calling for the resignation of Kourilsky and board chairman Michel Bon.

In these circumstances, says Nadine Peyrolo, a Pasteur spokeswoman, the board felt it no longer had the legitimacy needed to govern the institute. A new board that has the confidence of staff will have renewed legitimacy, suggests Sophie Chevallon, a spokeswoman for the science ministry, which has a permanent seat on the board. "We hope this will help to find a way out of the crisis," she says.

An extraordinary meeting of the assembly has been set for 15 March to elect a new board. One of its most urgent tasks will be to re-elect Kourilsky for another six-year term — or replace him.

Shortly after the board resigned on 12 January, the Pasteur's management e-mailed staff to announce that Kourilsky intends to send them all a letter detailing his actions and informing them of his "desire to open up a dialogue" with them.

But in a telling sign of the climate on campus, an internal memo from the institute's management says the dialogue will take the form of an Internet chat between staff and Kourilsky on 21 January, in which staff will have "perfect anonymity" so that "everyone can express themselves freely".

Science lobby urges UK to divert funds from military fields

Philip Ball, London

Public funding of science and technology in Britain is too focused on weapons-based research. So claims “Soldiers in the Laboratory”, a report released this week by Scientists for Global Responsibility (SGR), a lobby group backed by some of Britain’s best-known researchers.

The report, written by Chris Langley, a neurobiologist with the Hertfordshire-based consultancy ScienceSources, asserts that up to half of British public spending on military research and development should be diverted to more socially useful activities. It recommends spheres such as land-mine detection, conflict resolution, and water management.

SGR, a group of 600 scientists whose supporters include physicist Stephen Hawking and astronomer royal Martin Rees, argues in its report that such a shift would benefit both national security and economic competitiveness. It says that security would be better served by addressing global poverty issues, and that some British engineering companies would be fitter if they had to compete in non-military markets.

According to Stuart Parkinson, an electronic engineer and the group’s director, Britain spends 31% of its research and development budget on military work, a proportion that is exceeded only by the United States.

The report argues that “the military sector has a disproportionate effect on science, engineering and technology”. Links between the military and the academic world are increasing, it says, citing the 2002 launch of the Defence Technology Centres — collaborations between industry, government and universities who work on defence problems. Three such centres exist so far, involving 18 British universities. And British companies with major interests in military research, such as BAE Systems and Rolls-Royce, currently fund some academic posts.

In a statement, the UK Ministry of Defence said that its weapons research is geared towards making weapons more accurate and bringing fighting to a “swift conclusion”, reducing civilian casualties.

Parkinson is not sure that the UK government will heed the report’s advice. But he feels that the time is right to raise such issues. “Since the end of the cold war, discussion around military research and development has almost disappeared,” he says. “We’d like to reopen that debate.” ■



Bombs away? The United States continues to run a fleet of aeroplanes that can carry nuclear devices.

Antinuclear groups push to keep treaty review in the air

Michael Hopkins, London

The Nuclear Non-Proliferation Treaty (NPT) will become marginalized unless governments take a treaty review conference seriously, antiproliferation groups are warning.

An impasse exists, they say, between the United States, which critics regard as unenthusiastic about the treaty, and nations that have no nuclear weapons but want to see signs of disarmament from those that do. This threatens to let the treaty wither on the vine, the groups argue.

A report released on 11 January by the British American Security Information Council (BASIC) and the Oxford Research Group (ORG) says the New York conference, which is scheduled for 2–27 May, is crucial to the treaty’s continued relevance. The two groups are planning a campaign in the run-up to the meeting to make governments take the conference more seriously.

“We could break up on 27 May with the NPT in disarray,” warns Ian Davis, director of BASIC. He says that “resentment and retrenchment” are brewing among the treaty’s signatories as a result of differing interpretations of it and a widespread perception that it favours established nuclear powers. The treaty, which came into force in 1970, calls on states with nuclear weapons to take concrete steps towards getting rid of them — but none has shown any sign of doing so.

The treaty’s relevance is also threatened by the fact that India and Pakistan, which have each tested nuclear weapons, and Israel, which is widely assumed to possess them, have declined to sign it.

Proliferation experts continue, nonetheless, to view the treaty as important,

because it is the main international agreement that seeks to restrain the proliferation of nuclear weapons.

The two groups plan to produce a series of non-technical research reports between now and May to detail the areas where progress might be made, says ORG director John Sloboda. They also aim to meet with key delegates, including Sérgio de Queiroz Duarte, the conference president and Brazil’s ambassador-at-large for disarmament affairs.

Such efforts are welcome, says Trevor Findlay, director of the London-based organization VERTIC, which promotes effective verification of nations’ compliance with agreements such as the NPT. However, he questions the claim by ORG and BASIC that the treaty is in danger of collapse. “I think it’s a longer-term danger,” he says. “Nuclear-weapons states have tended to ignore their disarmament obligations. But the treaty is vital to them and they know it.”

At the last NPT review conference, held in New York in 2000, nuclear states agreed on a 13-step programme to move towards global disarmament. But the United States, in particular, has reneged on parts of this deal, claims Matt Martin, a BASIC analyst based in Washington DC. He cites the Senate’s failure to ratify the Comprehensive Test Ban Treaty, a move that is called for as one of the 13 steps.

Martin says he is encouraged by last November’s decision by the US Congress to block funding for several new nuclear programmes, including one to develop ‘bunker-buster’ bombs (see *Nature* 432, 542–543; 2004). But his group wants to see signs from the nuclear-weapons states that they are prepared to work towards disarmament. ■

Brain-scan ethics come under the spotlight

Erika Check, Washington

Suppose you're a neuroscientist studying human depression. In one study, you take pictures of depressed children's brains using a highly accurate imaging technique. Examining the data one day, you discover that one of your subjects is missing his entire parietal lobe, a crucial area for interpreting visual information. What do you do next?

Most scientists would refer the image to a physician, but this poses risks for both researchers and their subjects. Many scientists are unsure how to handle the legal, ethical and practical questions of accidental findings that could have implications for their subjects' health. These 'incidental findings' occur in many areas of biology, but cause particular problems in brain imaging.

The techniques used to study the brain, such as functional magnetic resonance imaging, are extremely powerful and are used more and more. And incidental findings are common. In a study released last October, 82% of brain-imaging researchers said they had turned up such findings, and 2–8% of research subjects have "clinically significant" findings, such as tumours, malformations or serious disease (J. Illes *et al.* *J. Magn. Reson. Imag.* **20**, 743–747; 2004).

But handling incidental findings is complicated, especially in the United States, and procedures for dealing with them vary from institute to institute. "We are uncovering increasing rates of incidental findings and discovering that there is no uniform protocol for dealing with them," says Judy Illes, an imaging researcher who directs the programme in neuroethics at the Stanford Center for Biomedical Ethics in California.

On 6 and 7 January, Illes led a workshop at the National Institutes of Health (NIH) in Bethesda, Maryland, to fill what one participant described as "a policy vacuum" on the question. About 50 scientists, physicians, lawyers and ethicists

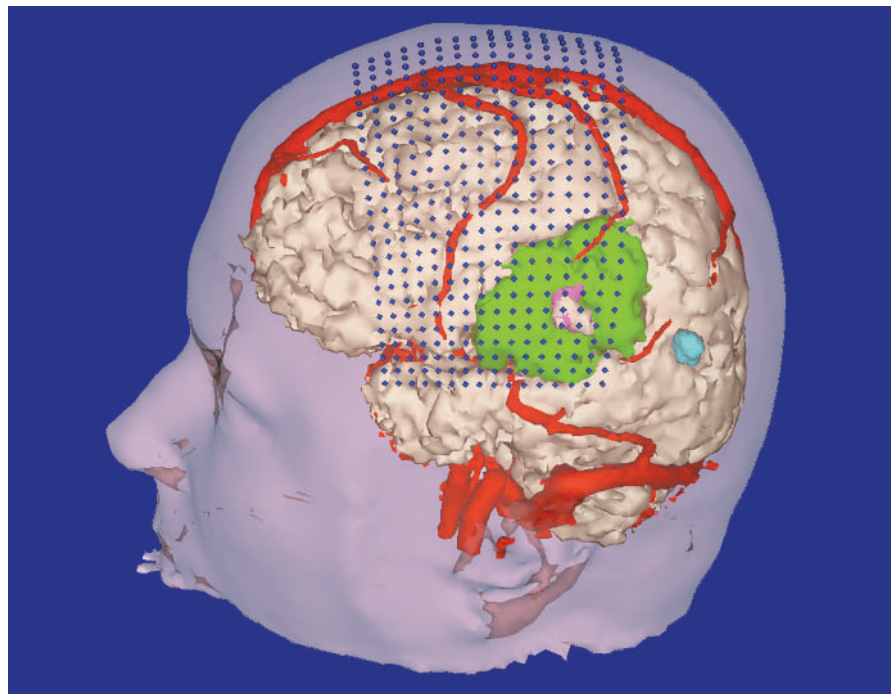


Judy Illes is working on ethical guidelines for scans.

attended and agreed that every researcher should plan to cope with incidental findings before beginning brain-imaging studies.

Most felt that researchers should be allowed to draw up their own plans for dealing with these findings, but many were uncomfortable with the idea that basic researchers should ever be expected to spot medical information in non-clinical studies.

"It's changing the responsibility of a principal investigator, who has no medical training, to that of someone who is diagnosing a disease," said Steven Grant, a neuroscientist at the National Institute on Drug Abuse,



Three-dimensional brain scans, such as this magnetic resonance image, can reveal nasty surprises.

based in Bethesda, Maryland. "That changes the relationship."

For some, the best way to solve the problem is to ask a trained medical professional to examine every scan in every brain-imaging study. David Yousem, a neuroradiologist at Johns Hopkins University in Baltimore, Maryland, supports this idea: "It's about an ethical obligation to the person you recruited to the study to inform them of a medical abnormality," he said.

In the study by Illes *et al.*, 23% of researchers surveyed followed this policy. The NIH goes one step further by requiring its on-campus investigators to perform a clinical scan of every research subject, in addition to any research scans. But workshop participants agreed that this was not a good idea because it can expose people to unnecessary risks from extra procedures.

Most researchers opt instead to send a radiologist only the brain scans that look obviously suspicious, such as the scan missing an entire parietal lobe. Illes *et al.* found that 53% of researchers they surveyed did this.

But a case-by-case referral to a radiologist raises a host of other problems. An incidental medical finding may affect a subject's ability to buy health insurance, for example, and findings that turn out to be medically insignificant can still cause tremendous stress for research subjects and their families.

Walter Schneider, a psychologist at the University of Pittsburgh, Pennsylvania, said that a participant in one of his research studies decided not to go to graduate school because of a perceived defect in a brain scan.

For some basic researchers, the scariest aspect of entering the medical realm is the threat of lawsuits. If they accept any responsibility for detecting medical information, researchers fear, they will be legally accountable for diagnosing brain problems that are completely unrelated to their studies. "One of the things that should be taken into consideration is the clinical research that won't be done because of the cost, and because the liability makes people nervous," said Alan Evans, a neurologist at McGill University in Montreal, Canada.

Lawyer Susan Wolf of the University of Minnesota, Minneapolis, pointed out that US courts are becoming increasingly likely to hold researchers responsible for their subjects' welfare. In Maryland, for example, a state appeals court decided in 2001 that researchers studying lead abatement procedures in Baltimore should have warned some children in the study that they had dangerous levels of lead in their blood. "We are already out of the era in which researchers had no legal responsibilities to their subjects," Wolf said.

Illes had hoped to agree general guidelines at the January meeting. But after heated debate, the group could not even decide whether every study should refer suspicious findings to a medical professional. Illes hopes to convene additional meetings to forge agreement on this point, among others. ■

Paediatrician launches bid to become Iran's next president

Paris A prominent Iranian scientist is set to stand in the nation's presidential elections in June. Mostafa Moin, who was widely respected as science minister, was chosen in late December as the candidate of the reformist party, the Islamic Iran Participation Front (IIPF).

Iran's president, Mohammad Khatami, cannot be re-elected as he has already held the two terms of office allowed under the constitution. Moin, a paediatrician and medical researcher, is president of the Immunology, Asthma and Allergy Research Institute in Tehran. He was science minister from 2000 to 2003, when he resigned, purportedly in protest at conservative policies and mass arrests of students.

Although his candidacy might yet be disqualified by the Guardians Council, a



Mostafa Moin hopes to win Iran's June election.

constitutional oversight body, Moin is already setting out some policies, saying that he favours promoting academic freedom. He also wants to develop Iran's economy through science and technology, and says he hopes to reverse the brain drain of the country's best scientists.

Inventor of blue LED sees red over settlement

Tokyo A record-breaking award of ¥20 billion (US\$195 million) made to a Japanese inventor by a Tokyo district court has been slashed in the court of appeal.

Shuji Nakamura, who is based at the University of California, Santa Barbara, reluctantly accepted a settlement of ¥840 million on 11 January from his former employer, Nichia Corporation of Shikoku. The award is for his pioneering work on technology to produce blue and green light-emitting diodes (LEDs) and blue lasers.

Nakamura said he was dissatisfied with the outcome but followed legal advice and settled during Nichia's appeal.

Last January, the Tokyo district court estimated that Nakamura was responsible for

Climate-change model becomes a class act

Washington To help take the mystery out of climate change, NASA last week released software that lets students do climate modelling in the classroom. The Educational Global Climate Model, or EdGCM, developed at the agency's Goddard Institute for Space Studies in New York, can be downloaded free from the Internet.

Letting high-school and college students play at climate modelling "exposes the model's strengths and weaknesses in a way that scientific papers and newspaper articles frequently obscure", says project leader Mark Chandler, a climatologist at Columbia University in New York.

Based on a more complex model at Goddard, the software allows students to explore the impact on climate of many variables, such as



how much sunlight is reflected by the atmosphere. And, at the click of a mouse, it lets them do something that has so far proved beyond the world's industrial nations: they can cut atmospheric carbon dioxide levels in half.

♦ www.edgcm.org

half the value of patents covering a method for making the LEDs (see *Nature* 427, 478; 2004). But during the appeal, the upper court reduced the estimate of Nakamura's contribution after it took into account the risk taken by Nichia that the research might not work, and the role of other employees.

Bioethics law restarts Korea's stem-cell work

Tokyo The Korean scientist who leapt to fame last February for deriving stem cells from cloned human embryos is back in the lab.

Despite his work being hailed as a major step towards using stem cells to treat disease, Woo Suk Hwang's research has been on hold pending approval under a new bioethics law in South Korea. That law came into effect on 1 January, and 12 days later Hwang and his team at Seoul National University received authorization from the government to continue their research.

Hwang's initial success had been clouded by questions over whether recruitment of women volunteers for egg donation followed ethical guidelines, and allegations that he had used eggs donated by his graduate student (see *Nature* 429, 3; 2004). Hwang denied any wrong-doing but said that he would suspend his research until the bioethics law came into effect. Bioethicists hope that two national and ministerial committees created under the law will keep a closer eye on ethical issues in the future.

Search is on to find a verse to best sum up the Universe

London The best equations are said to have the expressiveness of poetry. So it is perhaps fitting that a competition launched by the British Association for the Advancement of Science on 12 January offers entrants the chance to pen a poem celebrating the themes of time, space and energy.

The competition calls on "poetic physicists

and physical poets of all ages" to submit a poem on the central themes of Albert Einstein's theoretical work. The competition marks Einstein Year — the centenary of three of his most influential papers — and winners will be announced during UK National Science Week on 11–20 March.

Famous entrants include astrobiologist Paul Davies ("Einstein came down like a wolf on the fold/Determined to slaughter the theories of old"), novelist Terry Pratchett ("Away beyond the speed of light/I'll write a novel in one night") and scientists Dr Bunsen Honeydew and Beaker from cult TV programme *The Muppet Show* ("Energy's simple, you just make an 'E'/ By timesing an 'M' and squaring a 'C'").

If you think you can do better, visit www.the-ba.net/universe and have a go.

Government guidelines weigh in over obesity

Washington Obesity has joined problems such as cardiovascular disease as a central issue in the US administration's advice on healthy eating. Released on 12 January, *Dietary Guidelines for Americans* highlights the need to reduce total calorie intake.

The guidelines, published every five years by the health and agriculture departments, also advise eating more fruit and vegetables, and minimizing intake of refined sugars.

"They look like the strongest dietary guidelines yet produced," says Michael Jacobson, who heads the Center for Science in the Public Interest, a nutrition advocacy group in Washington DC.

But Jacobson and other experts warn that the public will ignore the guidelines unless the government promotes them more vigorously. They argue that the fight against obesity requires education campaigns as well as new legislation, for example to subsidize healthy foods and curb food advertising aimed at children.

♦ www.healthierus.gov/dietaryguidelines

All pain, no gain?

Exercise is good for you, or so we always thought. But, as Alison Abbott learns, your genes don't always cooperate.

M. POWELL/GETTY IMAGES

When Claude Bouchard set out to see whether genes play a role in physical fitness, he assumed, like most people, that exercise training makes everyone fitter. Although he expected genes to modulate some individual responses to diet and exercise, he also anticipated that regular workouts would improve fitness indicators such as lung efficiency and blood cholesterol for everybody.

Some 20 years later, it has become clear from the work of Bouchard and others that this is not the case. Looking at certain measures of fitness, some people actually fare worse after exercise, whereas others show little or no improvement.

But this isn't vindication for couch potatoes. Everyone's health improves in some way or other from exercise, but just how it improves is largely dependent on genes. Now, the growing field of fitness genetics is attempting to tease those genetic components apart, and the studies are generating fresh insights into the benefits of exercise as well as unexpected pay-offs for medicine.

Bouchard's attempts to track fitness genes began in the mid-1980s at Laval University in Quebec, Canada. He and his colleagues focused on the maximum amount of oxygen absorbed by the body from a lungful of air — a standard measure of aerobic fitness, usually abbreviated as $V_{O_{2max}}$. They found that most people can get more oxygen out of each breath after training but that a minority were

no better off, regardless of how efficient their lungs were at the start. Because the variation was much less extreme within pairs of identical twins, Bouchard concluded that the effect was largely dictated by genes¹.

That initial study was fairly small, so Bouchard extended the work in 1992 by helping to set up a multicentre research effort called the HERITAGE Family Study, which is still running today. Now based — together with Bouchard — at the Pennington Biomedical Research Center in Baton Rouge, Louisiana, the study's main data set comes from some 740 sedentary adults who were subjected to an intense exercise regime in the lab. The researchers monitored changes in the participants' blood pressure, heart rate, blood chemistry and $V_{O_{2max}}$ over 20 weeks.

Survival of the fittest

The study's main aim was to determine how exercise reduces risk factors for cardiovascular disease and diabetes, but Bouchard and researchers at the four other collaborating institutions also took blood samples for genetic analysis. "We were trying to find as many genes as possible that influence fitness and performance," Bouchard says.

The resulting reams of data and frozen blood samples are still being analysed, but the results so far confirm Bouchard's earlier stud-

ies. The average increase in $V_{O_{2max}}$ after the training programme was 19%. But 5% of the subjects had virtually no change, and another 5% had improved by more than twice the average amount. Similarly, most people had lower exercising heart rates and blood pressure after the training programme — an indication of improved fitness — but the extent of the reduction was extremely variable. In a few people there was even a small rise in these numbers².

Much of this variability seems to be attributable to genes. The researchers found more variation between than within families, suggesting at least a portion of a person's ability to benefit from exercise is inherited. "We concluded that just about half of the difference in trainability was heritable," says Tuomo Rankinen, the study's project manager.

It is unclear to what extent fitness parameters such as $V_{O_{2max}}$ are indicative of long-term health prospects, but even presumed health indicators such as cholesterol, a factor in heart disease, did not follow the expected pattern of more exercise is better. Conventional wisdom has it that regular exercise reduces the risk of heart disease by raising blood levels of high-density lipoprotein (HDL) cholesterol, a complex that helps prevent cholesterol from forming fatty deposits on blood-vessel walls. This is considered one

"Our work on athletes is feeding back into the clinic. How efficiently we use oxygen is decisive when we are desperately sick."

— Hugh Montgomery



those who exercise regularly, Rankinen says.

One way to begin to untangle these apparently contradictory effects is to go after the genes involved. This could ultimately reveal a great deal about how exercise produces health benefits, and may lead to treatments for diseases of metabolism and physiology.

To this end, scientists at the HERITAGE study are scanning the genomes of participants for gene variants that occur more frequently in association with different fitness responses. Although some metabolism genes have been identified that may play a role, the most strongly linked gene so far is *Titin*. This produces protein fibres that contribute to the elasticity of heart muscle cells. It may be that some forms of the gene allow the heart to pump larger volumes of blood than others³.

Physical attractions

Other teams are also on the hunt for fitness genes. Paul Williams, a health researcher at the Lawrence Berkeley National Laboratory in Berkeley, California, for instance, suspects that a gene related to the synthesis of HDL cholesterol might be involved. Ten years ago, he found that people who have an easier time taking up running after leading sedentary lives also started out with higher levels of HDL cholesterol in their blood — and increased those levels more quickly — than those who find running difficult⁴. It turns out that an enzyme that boosts HDL cholesterol is found in ‘slow-twitch’ muscle fibre, the type that takes longer to fatigue and so makes distance running easier. Williams is now beginning a large-scale genetic study to see whether differences in that enzyme are associated with differences in lifestyle choice.

Meanwhile, Gaston Beunen, a sports scientist at the Catholic University of Leuven in Belgium, is looking at the half-dozen or so key genes that contribute to the synthesis of myostatin, a protein that blocks new muscle growth. His study of some 300 young sibling males, published in May last year, hints that three of these genes may help to determine a person’s physical strength⁵.

In the end, the number of fitness-linked genes is expected to be large. So far, more than 100 appear in the literature, most of which have been identified in the past four years⁶, although in many cases more work is needed to confirm the link. And some of these now seem likely to prove their worth in the clinic.

One gene drawing a lot of attention encodes an enzyme called ACE, or angio-

tensin-converting enzyme. ACE activates the hormone angiotensin, which helps to maintain blood pressure and promotes the growth of the heart in response to exercise. One common gene variant, known as *ACE D*, makes more ACE than the other common version, *ACE I*. And athletes who have inherited *ACE D* from both parents experience about three times more heart growth in response to exercise than those who have inherited two *ACE I* genes⁷. They also seem to perform better in sports that rely on sheer strength and power, such as weight-lifting or sprinting. The *I* variant, in contrast, is more common among elite athletes in endurance sports such as long-distance running and swimming, which require more efficient metabolic use of energy and oxygen^{8,9}.

As the lower levels of ACE associated with *ACE I* improve endurance, Hugh Montgomery, a cardiovascular geneticist at University College London, wondered whether *ACE I* might also be advantageous to those suffering serious illness. He found that children with potentially deadly meningitis were more likely to require intensive care or to die if they had two copies of the *ACE D* gene rather than two copies of *ACE I*¹⁰. His team also found that premature babies with *ACE I* fare better¹¹. “Our work on athletes is feeding back into the clinic,” says Montgomery. “How efficiently we use oxygen is decisive when we are desperately sick.”

It may eventually be possible to help such patients with drugs that slow down ACE activity. Already, in unpublished work, ACE inhibitors have been shown to reduce muscle wasting in mice. And London-based drug company Ark Therapeutics is currently running final-stage clinical trials on the use of the ACE inhibitor imidapril to treat severe muscle wasting in cancer patients.

Fitness genetics may be feeding ideas into the clinic, but could genetic destiny become a new excuse for couch potatoes? “If they think their performance is limited by their genes, people tend to give up,” says Montgomery. “People are afraid of trying and failing — it’s part of the human condition.” Nevertheless, his advice to those who long to be fitter is to do serious exercise come what may. For everyone, it seems, there is at least some benefit. ■

Alison Abbott is Nature’s senior European correspondent.

1. Bouchard, C., Dionne, F. T., Simoneau, J. A. & Boulay, M. R. *Exerc. Sport Sci. Rev.* **20**, 27–58 (1992).
2. Bouchard, C. & Rankinen, T. *Med. Sci. Sports Exerc.* **33** (Suppl.), S446–S451 (2001).
3. Rankinen, T. et al. *Physiol. Genom.* **15**, 27–33 (2003).
4. Williams, P. T., Stefanick, M. L., Vranizan, K. M. & Wood, P. D. *Metabolism* **43**, 917–924 (1994).
5. Huygens, W. et al. *Physiol. Genom.* **17**, 264–279 (2004).
6. Perusse, L. et al. *Med. Sci. Sports Exerc.* **35**, 1248–1264 (2003).
7. Myerson, S. G. et al. *Circulation* **103**, 226–230 (2001).
8. Nazarov, I. B. et al. *Eur. J. Hum. Genet.* **9**, 797–801 (2001).
9. Tsianos, G. et al. *Eur. J. Appl. Physiol.* **92**, 360–362 (2004).
10. Harding, D. et al. *Am. J. Respir. Crit. Care Med.* **165**, 1103–1106 (2002).
11. Harding, D. et al. *J. Pediatr.* **143**, 746–749 (2003).

of the key benefits of taking up sports such as running. But the HERITAGE data show that training does not inevitably increase levels of HDL cholesterol. In fact, in about one-third of exercisers, the level of the complex fell.

Does this all mean that exercise could actually be bad for those of us with the ‘wrong genes’? Not at all, insists Rankinen. “We found not a single ‘universal non-responder,’” he says. In other words, everyone improved on some score. Even those who could not raise their $\dot{V}_{O_{2max}}$ through exercise were still getting some other health benefit such as higher HDL cholesterol levels or lower blood pressure. And overall, the HERITAGE data show that the risk of cardiovascular disease and type 2 diabetes falls in

The premier division

Since he took over as Harvard president in 2001, Larry Summers' style and vision have divided the university. As his plans for expansion step up a gear, Summers tells Helen Pearson why it is time for Cambridge to face up to the need for change.

Two faces of Harvard University sit on opposite banks of the Charles River, and only one of them looks pretty. The attractive side is what you might expect from the oldest and most venerable seat of learning in the United States: weathered brick façades, hushed rooms housing precious intellects, and rows of bicycles to ferry students to the some of the world's most hallowed lecture halls.

The other face of Harvard is one with which few outsiders are familiar, and is, for now, a comparative wasteland. A motley collection of industrial plots, rail yards, disused warehouses and a second-hand car lot make up this side of the university, in the distinctly unprecious Boston suburb of Allston.

It takes considerable imagination to envisage the wilderness of this second site being transformed into an academic hub as vibrant as its historic neighbour. But soon after Larry Summers became Harvard's president in 2001, that was the plan he put forward. Summers, a brilliant economist, gained national prominence as treasury secretary and economic adviser to President Bill Clinton. Under Summers' plan, Harvard intends to spend billions of dollars turning Allston into a campus whose size will eventually exceed that of the university's traditional base in Cambridge.

All change

The Allston plan is just part of Summers' mission to overhaul the university. He also wants to move science centre-stage in an institution that, although diverse, has traditionally enjoyed particular pre-eminence in the humanities. He plans to revamp undergraduate education — and, contentiously, he wants the central administration to exert more direct influence over the university's traditionally autonomous faculties.

Summers' direct style and sweeping agenda have, unsurprisingly, managed to irk

some Harvard academics. Although their criticism has mostly been internal and low-key, you don't have to wander the corridors there very long to find it. The critics object to the substance of Summers' plans, the way in which he has introduced them, and sometimes even to the man himself. "Some members of faculty dislike him intensely," says one critic, particle physicist Gary Feldman.



Summers understands these criticisms but contends that his plans' benefits will far outweigh any costs. The university, he says, is obliged to embrace the kinds of changes he has introduced if it is to retain its position as one of the world's pre-eminent academic institutions. "It worries me much more if a project has no enthusiastic champions than it does if it has a certain number of energetic detractors," he says.

With a history, assets and a reputation as rich as Harvard's, there is plenty to lose if

Summers is wrong. Founded in 1636, the university is the oldest higher-education institution in the United States. It is probably the wealthiest university in the world, boasting an endowment of more than \$22 billion.

On the academic front, Harvard also flaunts impressive credentials: it shares the number one spot with Princeton in the widely quoted *US News & World Report* rankings of US colleges for 2005, and at the last available count, boasted 156 members of the prestigious National Academy of





Near wastelands in Allston (left) are set to rival Harvard's traditional Cambridge home (above).

Sciences — more than any other institution. Harvard's traditional position at the apex of the US university system also ensures that others tend to follow its lead on questions such as how undergraduate courses should be structured.

Summers' management challenge at Harvard is magnified by the university's strongly decentralized structure. Its roughly 2,400-strong professorial staff is split into a loose federation of nine faculties. Most of these, including the Faculty of Arts and Sciences and Harvard Law School, tend to be clustered around historic Harvard Yard in Cambridge. A second campus, in the larger, neighbouring city of Boston, houses Harvard Medical School and the School of Public Health, and the third campus in Allston is already home to Harvard Business School.

The bottom line

A quaint Harvard expression — “every tub on its own bottom” — is sometimes used to describe the large degree of autonomy that each faculty enjoys. The faculties maintain considerable control over their own affairs and budgets. Traditionally, the president's main activities have been raising funds externally, while offering gentle guidance to the powerful faculty deans. Summers' predecessor, Neil Rudenstine, is said by many staff to have fitted that mould.

Summers stepped into the position in July 2001, when a colossal fund-raising drive by Rudenstine, as well as healthy growth of the endowment, had put the university in a particularly strong financial position. It was already poised to expand physically, after several years spent buying up wide swathes of land in Allston, but the university lacked concrete plans for how this should be done.

The Harvard Corporation, the executive board that appointed Summers, clearly anticipated that his background would help him to make an impact at the university. After building his reputation as an economist at both Harvard and the Massachusetts Institute of Technology, Summers was appointed chief economist for the World Bank in 1991. Under Clinton, who was elected in 1992 largely on a pledge to fix the US economy, Summers came to the US Department of the Treasury, where he became secretary — and the president's chief economic adviser — in 1999.

Summers shook things up at Harvard almost from the day he took over as president. In his inaugural address, he laid out plans to reform the university's undergraduate education, its decentralized structure and scientific research. By the end of 2001, he was engaged in a public row with Harvard's

best-known African-American academic, the philosopher Cornel West, who subsequently resigned and moved to Princeton University in New Jersey.

Encounters such as that one soon earned Summers a reputation among the staff as being opinionated, overly dismissive of their concerns, and even domineering. As one professor puts it, Summers is used to thinking of himself as the brightest person in the room. But in meetings at Harvard, he's likely to be dealing with people who have the same view of themselves.

Shaping history

In an interview last month at the New York Harvard Club, a hushed Manhattan building clad in the university's hallmark shade of crimson, Summers sketched out the key elements of his vision for the university.

A major goal, he says, is for Harvard to invest more heavily in science, particularly in interdisciplinary research. “I believe that when the history of this period is written 250 years from now, what happens in the life sciences and technology during the next quarter-century is likely to be a large part of it,” he says. Without such an approach, he argues, Harvard risks losing its leadership position. “One of the ways in which some of the British universities have lost positions of pre-eminence is by inadequate investment in science,” he notes.

As part of this commitment, one of Summers' first steps in 2001 was to hire the then director of the National Institute of Mental Health, Steven Hyman, to be provost — in effect, his top administrator. The two have already overseen several new science initiatives, such as the establishment of the Harvard Stem Cell Institute — an effort to bring together researchers studying embryonic and adult stem cells — and a department of systems biology, which seeks to understand entire biological systems by analysing large data sets.

Summers' ambitions for science are also evident in a vast, multibillion-dollar fund-raising drive just beginning at the university. Those involved say the details are undecided and that, at this stage, they are approaching key donors to build up the capital before making a public announcement. But they say that the campaign, which is likely to be seeking at least \$4 billion, will probably be the largest the university has ever undertaken — and will be distinguished from those in the past by the large proportion of money devoted to science projects.

To identify the scientific projects that Harvard should invest in, a 15-member task force, established in 2003 and led by Hyman, asked staff for their proposals. The group whittled down the 70 submitted to an initial shortlist of

“The very substantial physical resources we have available in Allston will allow us to define the Harvard of the twenty-first century.”

— Larry Summers

13. These include a project to understand the origins of the life in the Universe involving biologists, chemists and astrophysicists, and one to tackle global health, which would bring together experts in areas such as infectious disease, mental health, diagnostics, policy, economics, law and bioethics.

A second goal is to reform undergraduate education. Summers thinks that all undergraduates need to gain a more extensive grounding in modern scientific subjects such as genomics, so that they emerge as well versed in science as they are in the humanities. He also wants to see them all spend some time studying abroad. Both priorities are reflected in an undergraduate curriculum review currently under way at the university.

A third intention of Summers, which he doesn't spell out quite so explicitly, is to transform the collection of traditionally autonomous faculties into a single, more coherent university. He has done this partly by introducing more scrutiny of each school's budget and tenure appointments. Summers argues that this kind of supervision can save the university money, ensure that academic standards are consistent across the institution, and help the university to address academic questions that fall beyond the scope of any individual faculty.

Foundation stones

All of these ambitions will, Summers believes, be furthered by the successful transformation of Allston into a vast, new academic centre. "The very substantial physical resources we have available in Allston will allow us to define the Harvard of the twenty-first century," he says.

In a letter to the Harvard community on 21 October 2003, Summers outlined his ideas for renovating the 200 fragmented acres at Allston. He expects new science and technology facilities to form the heart of the new campus, alongside a new home for the School of Public Health and the Graduate School of Education, new student housing and an array of shops, museums and other attractions to ensure that students and staff alike flock to the site.

The plans for Allston aren't yet complete. But the occupants of the first scientific labs there, probably including the new institute for stem-cell research, are expected to be announced in the first half of this year.

Many academics, particularly those whose favourite projects are likely to benefit directly, are enthused by Summers' plans, and have embraced the opportunity to consider cross-cutting fields. "No past president has given us the freedom to think in those terms before," says chemist Greg Verdine, who has been involved in an initiative that combines chemistry and biology to, for example, try to



No doubt: Larry Summers believes Harvard must change or lose its lead in the academic world.

adjust the action of biological molecules.

But Allston, which embodies much of Summers' broader schemes for the university, has also become a focal point for criticism — particularly among scientists at Harvard.

Problems flared up in the Faculty of Arts and Sciences in October 2003, when staff members were alerted to the release of Summers' letter in an e-mail that they received only a few hours before they were due to discuss it in a faculty meeting. In this meeting and subsequent ones, many staff members expressed doubts about the plan. "To say there is faculty concern and lack of enthusiasm is a massive understatement," says one staff member.

Critical point

Critics raise three main concerns. First, they say that Summers and the administration failed to discuss adequately the best course for the university's development, and that such consultations that did take place were purely cosmetic. "As far as I can tell, Larry and a bunch of other high-level officials met, and then Larry just decided what he wanted to say," says astrophysics professor Alyssa Goodman.

Second, critics question whether an aggressive physical expansion is either desirable or necessary. They argue that the university could fulfil many of its needs for interdisciplinary work by careful expansion and collaboration within Cambridge, and

point out that small departments can be intimate and advantageous for collaboration. "Being big doesn't necessarily mean being good," says Feldman.

Third, they are uneasy about the isolation of the new campus. Its location, perhaps a 20–30-minute commute from the Cambridge campus, could divide rather than unite the scientific community, they say, interfering with the interdisciplinary collaborations it is supposed to nurture. They also worry that it could isolate science from the university's heart at Harvard Yard. "I think creating a science ghetto is a bad idea," says Goodman.

Summers and Hyman are unmoved by the criticisms. They claim that initial dissension over the Allston plans has dissipated as members of staff have had the chance to participate in the planning process, and get fired up about the opportunities there. "If you'd listened to the discussion that took place at a recent faculty meeting, it had a relatively different tone from that at the meeting a year ago," Summers observes.

The two men also brush off the suggestion that expansion at Harvard is unnecessary. Growth and change, they say, is the only way to ensure that the university is able to fully engage in emerging areas of research 50 years from now. "The danger to Harvard is that people run the risk of becoming too comfortable and complacent because it is so successful," says Hyman.

And they urge the critics to engage in the renovation process, instead of opposing it.

"The danger to Harvard is that people run the risk of becoming too comfortable and complacent because it is so successful."

— Steven Hyman

"I've asked people to focus their energy on what they want to do and how I can help them do it — and not to focus as much energy on what they think other people should not be allowed to do," Summers says.

Whether academics choose to follow or fight Summers' advice, they do seem to be learning to live with his approach. Some say that they find his decisiveness refreshing — and preferable to the agonizingly slow and consultative process that sometimes dogged decision-making in the past. "Every time you make a big, bold decision you're going to piss people off — but presidents have to do that," says professor of psychology Marc Hauser.

Nevertheless, in a place that thrives on academic debate, Summers' combative approach will continue to send ripples of dissent through the quads and corridors. And Summers has yet to impress everyone that his way of working is suited to the venerable institution. "You don't run Harvard the way you run the Department of the Treasury," says history of science professor Everett Mendelsohn. ■

Helen Pearson works in New York for Nature's online news team.

Croatian minister rejects 'meddling' claim

Mediterranean Institute's link with university is intended to ensure academic freedom.

Sir—Your News story about the Mediterranean Institute for Life Sciences, "Croats protest that science minister is 'meddling' in MedILS" (*Nature* 432, 540; 2004), alleges that some researchers believe I am trying to stack the advisory board with faculty members from the University of Split. In fact, I am securing the highest level of academic freedom for MedILS within Croatian law and European tradition.

Faithful to the contract that the former Croatian government signed with Miroslav Radman, the current government will do its best to hand him the renovated building in Split in May 2005. The total cost of renovation, some €4.6 million (US\$6 million), has been appropriated from the Croatian budget. This is more than the capital investments in all 26 Croatian public scientific institutes for 2005. As

a result, the government is compelled to take a similarly large World Bank loan for the restructuring of the 700-strong Rugjer Boskovic Institute (RBI). But, unlike the unburdened MedILS, the RBI is assuming numerous obligations to the World Bank.

Some dissenters say that public funds should not be allocated to a project that has no scientific plan, no business plan, no mechanism for supervising the spending of public funds, no pledge to share future proceeds from intellectual property with taxpayers and no commitment by Radman to move to Split from his current base in Paris in order to manage MedILS.

To mitigate pressure to put MedILS under the control of my ministry, I have loosely associated MedILS with the University of Split. In compliance with the law, we formed a board of trustees: two

publicly appointed (by the university and the ministry) and three internal (MedILS-appointed). Affiliation with the university will assure the public with regard to the spending of public funds, guarantee the highest level of academic freedom and result in public trustees being academics rather than Zagreb bureaucrats. Trustees deal with administrative management and do not 'meddle' in scientific matters.

Croatian law requires appointment of a separate and independent scientific board to manage scientific affairs. This will consist of international scientists of Radman's choice. Further details are available on the ministry website, www.mzos.hr.

Dragan Primorac

*Ministry of Science, Education and Sports,
Trg hrvatskih velikana 6, HR-10000 Zagreb,
Republic of Croatia*

Insect collection ready to spread its wings

Sir—Your News story "Curators bugged by museum's vision for insect collection" (*Nature* 432, 659; 2004) gave the impression that Darwin Centre II (DCII) will be incompatible with a 'cyber-infrastructure' future for taxonomy.

Although we are certain that taxonomy and collections-based research in general will look very different ten years from now, no consensus exists as to the details of this future. Designing the best available work-spaces for such a rapidly evolving science is not easy. As scientists we are accustomed to loudly debating competing ideas; it is no different with regard to facility design.

The main issue is whether sufficient flexibility exists for research groups of varied sizes to have immediate access to specimens of entire taxa. This issue is being addressed in a way that balances the best possible long-term conditions for irreplaceable specimens against efficient access during bursts of intensive research.

Far from resisting change, the Natural History Museum is leading the transformation of taxonomy into a 'big science' capable of rapidly advancing our understanding of the natural world. Its insect collection is unique, as the only place where one can directly compare specimens of more than half of all known species. DCII will provide modern research laboratories, state-of-the-art conditions for collections and an opportunity to open this research to the public. Controversy over details should not cloud the big picture;

DCII will provide first-rate facilities for world-class taxonomic research.

Quentin D. Wheeler

*Department of Entomology, Natural History
Museum, Cromwell Road, London SW7 5BD, UK*

Alternative views of amphibian toe-clipping

Sir—In News and Views ("Ethics and amphibians" *Nature* 431, 403; 2004), Robert M. May discusses a study by M. A. McCarthy and K. M. Parris on the effects of toe-clipping on amphibians. This is a standard technique for uniquely marking animals in ecological research (see guidelines at www.asih.org/pubs/ASIH_HACC_Final.pdf). McCarthy and Parris show that return rates decreased as the number of toes clipped increased in four frog species; they and May raise ethical and practical questions about the technique. There is, however, a fuller picture to be considered.

Several studies have found no negative effects of toe-clipping. McCarthy and Parris suggest that this may be due to low statistical power, but this has not been shown in all cases in which no effect was observed (see, for example, J. J. van Gelder and H. Stribosch *Amphibia-Reptilia* 17, 169–174, 1996; and J. A. Ott and D. E. Scott *J. Herpetol.* 33, 344–348, 1999). Effects of toe-clipping evidently vary among species and so must be assessed accordingly.

Alternative marking techniques often cannot be used. Many adult frogs are less than 20 mm long, as are juveniles of larger species. The smallest passive integrated

transponder (PIT) tags are about 10 mm long, too large for small frogs. Many species cannot be marked using other approaches, such as elastomers, alphanumeric tags or freeze-branding, and do not have consistent, identifiable markings.

Other techniques may have worse effects, as M. Schlaepfer has shown (*Herpetol. Rev.* 29, 25–26; 1998). Implantation of PIT tags often involves surgery on the body cavity; even implantation under the skin carries a risk of more serious infection than in an extremity. Many ecological studies require trading the risks of a marking technique against the gains in understanding. More research evaluating these risks is needed.

Far from acting with 'casual barbarity', biologists who use the toe-clipping technique do so after considering alternative techniques and with the approval of institutional animal care and use committees. The resulting data are essential for managing threatened populations, as well as for other purposes: toe clippings are sources of DNA for genetic studies, and act as samples for identifying the diseases implicated in catastrophic amphibian declines.

We believe it is less ethical to sit back and watch species slip into extinction than it is to use the best available methods to help to conserve them.

W. Chris Funk*, Maureen A. Donnelly†, Karen R. Lips‡

**Section of Integrative Biology, University of Texas,
Austin, Texas 78712, USA*

*†Biological Sciences, Florida International
University, Miami, Florida 33199, USA*

*‡Department of Zoology, Southern Illinois
University, Carbondale, Illinois 62901, USA*

The Einstein chronicles

Two volumes of correspondence put Einstein's work in a historical context.

The Collected Papers of Albert Einstein, Volume 9. The Berlin Years: Correspondence, January 1919–April 1920

edited by Diana Kormos Buchwald, Robert Schulmann, József Illy, Daniel J. Kennefick & Tilman Sauer
Princeton University Press: 2004. 780 pp.
\$110, £72

The Born–Einstein Letters 1916–1955: Friendship, Politics and Physics in Uncertain Times

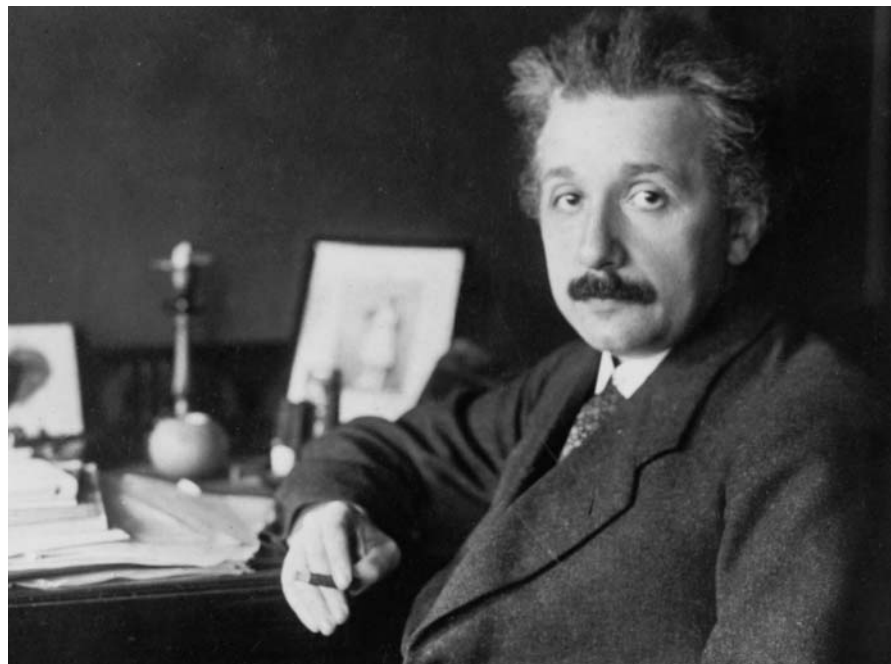
by Albert Einstein & Max Born
Macmillan Science: 2005. 256 pp.
£19.99, \$26.95

Gerald Holton

On 22 September 1919, Albert Einstein, then aged 40, received at his Berlin home a brief telegram that signalled that his life would shortly change, suddenly and forever. The telegram contained the news that Arthur Eddington had shown during a solar eclipse that starlight is deflected by the Sun's gravity — an effect predicted by Einstein's general theory of relativity. The scientific details, when published that November, made Einstein almost overnight the icon of charismatic scientific genius. The very stars had announced, to a public still in mourning after the barbaric First World War, that old conceptions were being overturned. Einstein's vast correspondence during the months that followed reveal the reactions, ranging from adoration to vicious attacks.

This tumultuous period is recorded in volume 9 of *The Collected Papers of Albert Einstein*. The 399 letters and other documents included, most of them previously unpublished, offer us an intimate view of the remarkable multidimensionality of Einstein's life. Much of the correspondence is of course about the scientific findings and controversies of the time, chiefly on the light-bending and red-shift tests of his general theory of relativity, the attempt to find common ground between quantum mechanics and relativity, and the programme to unify electromagnetism and gravitation.

Even an incomplete list of Einstein's scientific correspondents in this period will draw historians of science and other scholars to this volume: Paul Ehrenfest, Theodor Kaluza, Felix Klein, David Hilbert, Gustav Mie, Adriaan Fokker, Max von Laue, Hendrik Lorentz, Jean Perrin, Eddington, Willem de Sitter, Hedwig Kohn — and not forgetting the English physicist Robert W. Lawson, who persuaded Einstein to write an article for *Nature* in 1921 outlining his theory of relativity.



Man of letters: Einstein's correspondence provides a wealth of personal and scientific information.

Other letters deal with epistemological questions (such as exchanges with Moritz Schlick and Hans Vaihinger); with administrative, political and social problems, including Einstein's intensifying interest in a Jewish identity and in a form of Zionism. From more personal letters, we learn about Einstein's own precarious health in those days, and his deep grief as his mother is dying of cancer. Throughout there is evidence of his often difficult relations with his wife Mileva, who, soon after the family arrived in Berlin in 1914, returned to Switzerland with their two sons, for whom Einstein's letters show his great attachment. Also included are the decree of divorce from Mileva (dated 14 February 1919) and the certificate of his marriage to his cousin Elsa Löwenthal (2 June 1919, both registering their religion on the document as "Dissident").

There is as much as a page of editorial explanation and commentary on every name, place and allusion recorded in the corresponding letter — an astonishing combination of scholarship and detective work by the editorial team, now led by Diana Kormos Buchwald and continuing the tradition established by John Stachel, who edited earlier volumes. These annotations establish not only the relationship between different parts of the whole correspondence (the first volume was published in 1987), but also tell us much about the scientific and social milieu of the day, something only hinted at in the letters themselves.

For example, in a letter of November 1919 to Einstein, Lawson mentions in passing that he understands that "the conditions of life in Berlin are nearly unbearable". The editorial note explains: "The Allied blockade of Germany, which was extended past the end of the war to pressure Germany to accede to Allied demands, together with economic disruption from the war, caused a famine in central Europe in the aftermath of the war." There is then a discussion about the spread of disease.

Similarly, to clarify an obscure sentence in a letter by Fokker (dated 18 November 1919), the editorial note explains that, in testifying that month before the German official committee exploring the responsibilities of the military during the war, General Hindenburg "used the occasion to launch the 'stab-in-the-back' legend which blamed civilian groups such as the communists, Jews, and women for Germany's defeat in the war". The 'stab-in-the-back' legend "served as a catalyst for those embittered by Germany's defeat, fueling the consolidation of rightist and militarist attitudes in Weimar Germany".

All this carefully researched editorial commentary makes it easy to grasp the context and conditions underlying the correspondence. Einstein usually seems somehow lifted above it all, being saved in those dark days largely by what Max Planck pointedly identified as "*die freudige Sicherheit*" — the joyful certainty.

A 24-page introduction serves as a guide

to the main events and ideas in the correspondence; without it, it would be difficult to keep in mind the main themes when reading the chronological series of letters. There is also an explanation of the editors' triage of the voluminous correspondence: some letters are known but are inaccessible now, and the editors relegated other, less significant, letters to reference in a calendar at the end of the volume. And although the correspondence is presented only in the original language, mostly German, happily there exists a separately published English translation, provided by Ann and Klaus Hentschel (Princeton University Press, 2005).

Many of the most intimate letters in volume 9 are exchanges between Einstein and Max Born and his wife Hedwig. Here, Einstein can truly relax and share the good and the sad in his life, his scientific, social, musical and personal thoughts. Max Born, whose Nobel prize for his work in quantum mechanics was delayed until 1954, contributed to a wide spectrum of fields, including relativity, optics, crystal physics, epistemology, the theory of liquids and arms control. Max and Hedwig, who was a delightful companion and correspondent, left Germany for Edinburgh after Max was driven out from Göttingen in 1933. His son, Gustav, tells that his father had taken seriously Einstein's advice to leave immediately, "thus helping to save the family".

The 117 surviving letters between the Borns and Einstein from 1916 to 1955 were published as a book in German in 1969, with an introduction by Werner Heisenberg, at one time Born's assistant. It had a foreword by Bertrand Russell, who succinctly summarized what that book showed in its essence: "In an age of mediocrity and moral pygmies, their lives shine with an intense beauty. Something of this is reflected in their correspondence, and the world is richer for its publication." For most of the letters, Max Born provided perceptive editorial comments, explaining the circumstances of the time in which the letters were written.

An English translation of the book appeared in 1971, but has been long out of print. The new English-language edition of *The Born-Einstein Letters 1916-1955* — issued by Nature's publisher Macmillan as part of its new science imprint — is essentially unchanged from the earlier version, but greatly benefits from an extensive preface by Buchwald and the physicist Kip Thorne. To quote from the preface: "The letters

themselves constitute one of the most vivid and valuable testimonies to the development of modern science. They also tell us much about the personal hardships that Einstein and Born overcame during two world wars, the vagaries of academic life, the daily grind of administrative work, the steadfastness and frailty of human relationships."

Perhaps the most famous of these letters deal with the fundamental difference between these two good friends, with Born upholding what came to be the Copenhagen interpretation of quantum mechanics,

according to which the foundations of physics are laws of statistical nature. Einstein responded in

1926: "Quantum mechanics is certainly imposing.

But an inner voice tells me that it is not yet the real thing. The theory says a lot. But it does not really bring us any closer to the secrets of the 'old one'. I, at any rate, am convinced that He is not playing at dice."

This judgement, Born records, was for him "a hard blow", separating Einstein from the younger physicists of which Born, though only a few years younger than Einstein, considered him-

self part. But the friendship survived, not least by the intervention of Wolfgang Pauli, who cleared up a misunderstanding between Born and Einstein. The exchange of letters continued, testifying to their constant interests in developments ranging from atomic physics to cosmic rays, from quantum mechanics to stellar spectra, and of course relativity (the new 'ether drift' experiments, gravitational red shift, gravitational light deflection and gravitational waves).

But we hear also about the pair's intervention in humanitarian causes, the turmoil in politics throughout Europe after the First World War, programmes to help refugee scientists, and the devastating events of the Second World War. The new preface also contains valuable brief accounts of the way that physics, after the death of these two great minds, continued along lines they had pursued, showing in some detail how experimental and theoretical work in the past few decades has confirmed with great precision some of their daring speculations.

In short, these two related volumes act as a kind of microscope through which to view a turbulent period in the twentieth century.

Gerald Holton is Mallinckrodt research professor of physics and research professor of the history of science, Department of Physics, Harvard University, Cambridge, Massachusetts 02138, USA.



In touch: Max Born and his wife regularly wrote to Einstein.

Relativity revisited

The Meaning of Relativity (5th edn)

by Albert Einstein, with an introduction by Brian Greene

Princeton University Press: 2004. 192 pp.

\$14.95, £9.95

Special Relativity: A First Encounter

by Domenico Giulini

Oxford University Press: 2005. 160 pp.

£14.99, \$24.95

Einstein 1905: The Standard of Greatness

by John S. Rigden

Harvard University Press: 2005. 192 pp.

\$21.95, £14.95

Werner Israel

What, another three books on Einstein? At the last count on www.amazon.com there were 498 currently in print, and the proliferation of titles such as *The Private Albert Einstein*, *Einstein in Love* and *Einstein's Daughter* should ensure that no corner of his life is left untouched.

As a refreshing change, *The Meaning of Relativity*, *Special Relativity* and *Einstein 1905* deal exclusively with science. All three are valuable additions to the Einstein canon. *The Meaning of Relativity*, the master's own presentation based on lectures given at Princeton University in 1921, reappears in a reprinting of the final (1953) edition, which included as an appendix his parting thoughts on his last abortive bid to unify the gravitational and electromagnetic forces. This was in fact a revival, with modifications, of his earliest attempt, begun in 1925 but soon abandoned because, as he admitted in 1927: "As a result of numerous failures, I have come to the conclusion that this road does not lead us closer to the truth." It is curious that none of his later works mentions this early attempt. Not too much should be made of this; Einstein was always sparing with references (this entire book has just one). Yet one cannot help wondering: could it have slipped his mind that he had been down this path before?

Brian Greene's easy-to-read 24-page introduction touches on several developments since Einstein's time. One of these is superstring theory, which carries forward Einstein's quest for unification. Another is the recent discovery that the cosmic expansion is accelerating, with its strong implication that Einstein's self-styled "greatest blunder" — the cosmological constant — was perhaps not such a dumb idea after all.

Giulini's book *Special Relativity* has a narrower focus, concentrating exclusively on Einstein's special theory of relativity. It is particularly strong on experimental tests and ramifications of the theory, and on the evolution of relativistic ideas, from Galileo and Newton, through the nineteenth-century

Science in culture

The great portrait mystery

A disputed portrait of Robert Hooke may in fact show a contemporary.

Philip Ball

At the tricentenary of Robert Hooke's death two years ago, no one knew what he looked like. Despite Hooke's reputation as one of the principal architects of the scientific revolution, there were no known surviving portraits of him — an outcome, it was rumoured, of his enmity with Isaac Newton, who did all he could to erase Hooke's image after his death.

But in 2003, historian Lisa Jardine ruffled academic feathers by boldly claiming to have discovered a portrait of Hooke, which featured on the cover of her biography *The Curious Life of Robert Hooke* (HarperCollins, 2003). This painting has resided for more than a century in the Natural History Museum in London, where it was taken to be a portrait of the British naturalist John Ray (1627–1705) painted by the seventeenth-century artist Mary Beale. The painting was bequeathed as such to the museum in 1787 after the death of its former owner, the botanist William Watson.

Jardine argued that the visage looks nothing like other portraits of Ray, and that the features instead match some contemporary descriptions of Hooke, who was said to have bulging grey eyes and curly brown hair and to be of emaciated appearance. Others have found this evidence not only slender but also unconvincing: the face is certainly unusual, but does it really correspond in any regard to these accounts?

Now William Jensen, a specialist in the history of chemistry at the University of Cincinnati, has an alternative proposal. He points out in *Ambix* (51, 263; 2004) that the portrait can be superimposed remarkably well onto an engraving of another influential seventeenth-century scientist, the Flemish chemist and physician Jan Baptista van Helmont (1579–1644). The engraving appears in the 1648 edition of van Helmont's great work



Ortus medicinae, published posthumously by his son Franciscus Mercurius van Helmont (whose likeness is inserted behind his father's on the same page). Van Helmont's writings were never published in his lifetime because he was persecuted as a heretic by the Spanish Inquisition and forced to live under house arrest in Vilvoorde, near Brussels, until his death.

Particularly telling is the wispy moustache and underlip hair in the 1648 work, which is reproduced in the portrait thought to be of Ray. There is no record of Hooke having sported such facial hair. Hooke was only nine years old when van



Getting the brush-off: the above painting, thought to be of Robert Hooke, may have been based on an engraving of Jan Baptista van Helmont.

Helmont died, so there seems to be no possibility of the reverse confusion.

So where did the 'Ray' painting come from? Jensen says that Franciscus van Helmont had his own portrait made while he resided in England during the 1670s, and might have commissioned a picture of his father at the same time, based on the earlier engraving. But then who was the artist? And why did Watson, a hundred years later, have the false impression that he owned a painting of Ray by Mary Beale?

The haunting image clearly still holds mysteries. But the face of poor Robert Hooke may have vanished once more from history.

Philip Ball is a consultant editor for *Nature*.

aether theorists, and up to A. A. Michelson, Edward Morley, George FitzGerald and Hendrik Lorentz.

Have Lorentz and Henri Poincaré received less than their due in this great conceptual revolution? I think Giulini puts the case fairly: "In retrospect, special relativity seems palpably close in 1905, after all the preliminary works of Voigt, Hertz, FitzGerald, Lorentz, Larmor and Poincaré. But apparently it needed an unprejudiced newcomer to take the final step."

The mathematical demands of these two

volumes are not heavy (Giulini uses nothing beyond high-school algebra), but they do require close attention from the reader. In a lighter vein is John Rigden's enjoyable contribution, *Einstein 1905*. This is a month-by-month chronicle of 1905, Einstein's *annus mirabilis*, in which appeared in quick succession his four epoch-making papers on the photon hypothesis, brownian motion, special relativity and $E=mc^2$. Rigden explains the underlying ideas in clear, elegant, non-mathematical prose. Amusingly, of all of Einstein's 1905 works, the one most cited

today is none of the above (they are scarcely cited at all), but his PhD thesis on the determination of molecular dimensions. This is because the methods he used for it have been widely applied to such problems as the motion of sand particles in cement mixes and of aerosol particles in clouds. As Rigden remarks: "When a paper is so important that it could be cited in almost every paper, it is cited in almost no paper".

Werner Israel is in the Department of Physics, University of Victoria, Victoria, British Columbia, V8W 3P6, Canada.

A novel view of global warming

State of Fear

by Michael Crichton

HarperCollins: 2004. 603 pp. \$27.95, £17.99

Myles Allen

Jo Public is smarter than we think. When I reviewed the film *The Day After Tomorrow* for *Nature* (429, 347–348), I said that it couldn't really do any harm, and it made geophysics look cool, so we shouldn't worry about the liberties it took with thermodynamics. I was wrong. Surveys of public opinion conducted before and after the film was released found that it made people think climate change is less likely: "If that's what it's all about, there's no way..." Michael Crichton's latest blockbuster, *State of Fear*, is also on the theme of global warming and is, in its own way, equally likely to mislead the unwary.

A long plane journey to the 10th Conference of the Parties of the United Nations Framework Convention on Climate Change was the perfect excuse to read this book. "So," Crichton's hero (an MIT professor who moonlights as a secret-service agent) would interrupt to declare, "he's one of them." The central thesis of the book is that we scientists are collaborating with the environmental movement, bending facts in a cavalier manner to fit our mad global-warming theories — and when the facts won't bend far enough, we make them up.

It is a sad indictment of the image of modern science that many readers will find this thesis entirely plausible, even if they don't buy the specific scenario of eco-terrorists setting off a train of synthetic natural disasters to provide mood music for an international climate conference. Crichton himself clearly accepts it, and desperately wants his reader to do the same, to the extent that he keeps freezing the action to provide lengthy sermons, complete with graphics, about all that is wrong with the evidence for global warming. The ruse that Crichton uses to bring in these sermons is neat: a team of worried environmental lawyers, preparing a case against the US government, rehearse possible arguments that might be used by the defence.

Crichton has evidently been inspired by the eco-conspiracy theories of Bjørn Lomborg's *The Skeptical Environmentalist*. Both books demonstrate how intelligent reviewers, given a complex issue and sufficiently rich literature, can find support for whatever position they care to adopt. For example, Crichton opens with the fact that glaciers are advancing in Iceland and Norway, failing to mention that they are retreating just about everywhere else outside the polar ice caps.



Cold comfort: Argentina's Perito Moreno glacier is still advancing, but many others are in retreat.

Given the thousands of temperature records available on the Internet, it is unsurprising that he can find some to support the claim that much of the apparent warming during the twentieth century is due to urbanization.

Occasionally there is a straightforward error, such as the claim that the scientific community cannot explain why temperatures rose early in the twentieth century, flattened off and then rose again, while carbon dioxide levels were rising all the time — the explanation, a combination of natural and other anthropogenic drivers, has been available for years. And Crichton cannot resist bringing in a tsunami (presumably a precondition of the movie deal), although even he must know that no one has ever claimed a link between tsunamis and climate change. Although this is a work of fiction, Crichton's use of footnotes and appendices is clearly intended to give an impression of scientific authority. He appears to have succeeded, as the book has already been respectfully cited in the US Senate as a serious contribution to the climate-change debate.

Crichton does make some nice observations, for example about the outdoor gear worn to climate conferences by people who never stray from a computer (let's face it, we're badly paid office workers). With his medical background, he also finds the lack of double-blind studies in climate research troubling. Here he has a point, but the solution is far from obvious, given that we are all studying a single patient, and we learn in the first term of graduate school how the patient is believed to have behaved over the past century. Having contentious results re-analysed by independent groups is an excellent idea; the US National Science Foundation, to its credit, recently did just this with results from

the Microwave Sounding Unit. But Crichton is strangely silent on this one, presumably because the results were not to his liking.

Books such as *State of Fear* highlight the question of whether there is any such thing as a dispassionate scientific review. Is it just a matter of political taste whether to believe Crichton and Lomborg over, say, the Intergovernmental Panel on Climate Change (IPCC)? If the scientific community is to recover its standing in the world, we have to stand up to the pseudo-democratic vogue for treating everyone, regardless of expertise, as just another stakeholder. No science is infallible, but there is good science and there is bad science, and it is not just a matter of opinion which is which. A hallmark of good science must be the way it treats uncertainty. The IPCC said in 2001 that there is up to a one-in-three chance that the warming observed over the twentieth century might be "entirely natural in origin". If Crichton and Lomborg were equally frank about the chances of their basic premises turning out to be wrong, their scientific views would be a lot more credible.

But in the end, *State of Fear* is about people: the transformation of Peter Evans, a lawyer who drives a clean 'hybrid' car and is too shy to admit he has never fired a gun, into a gun-toting carbon-junkie at the wheel of a fuel-hungry sports utility vehicle. It definitely helps him get the girls. "There was something about him. Some surprising quality she hadn't noticed before." He had become... a climate sceptic.

That pretty much says it all about this book: Viagra for climate sceptics. ■

Myles Allen is in the Climate Dynamics Group, Department of Physics, University of Oxford, Parks Road, Oxford OX1 3PU, UK.

G. GALLARDO-TELAMAP

Schrödinger's mousetrap

Part 1: The trap is primed.

Ian Stewart

Rufus Jaeger looked out from the platform at a sea of faces. Every seat was full, and dozens of people were sitting on stairs and lining the walls. It made for an electrifying atmosphere, which was just how he liked it. The audience would remember this lecture for a long time.

Jaeger was head of the Quantum Optics Group at the University of Wentbridge, and he had been selected for a unique honour: to give a plenary lecture inaugurating the World Year of Physics. He was determined to give his audience an experience that would establish him as the leading authority in quantum physics. So this would not be any ordinary lecture. Jaeger was going to carry out his latest, revolutionary, experiment — the one he called Schrödinger's mousetrap — live on stage.

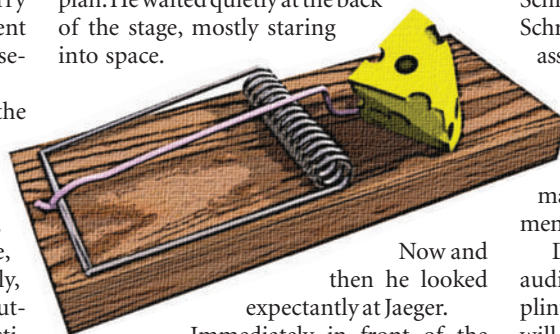
It had cost a small fortune to set the experiment up in the conference hall. Much of the stage was occupied by optical and other apparatus. Above, on a plinth, was a huge model of a mousetrap, complete with a wedge of Swiss cheese, cartoon-style with big holes. Fortunately, one of Jaeger's grants had money for 'out-reach' activities. The expenditure was justified by the presence of several journalists in the second row of the audience. Among them was Nigel Lorimer, a senior editor at *Nature*. He acknowledged Jaeger's presence with a wave, and Jaeger gave a brief nod in return.

Jaeger was an imposing figure: tall, well-built, with a flourishing beard that framed his face and sharpened his features. He could be charming when he decided that charm would get him what he wanted, and intimidating when a head-on confrontation suited his purposes better. He was sitting just left of centre stage, beside the apparatus. Next to him was his postdoc Ludmilla Shlomiuka, a strikingly attractive blonde with one of the best brains in the business. She was there to introduce Jaeger. She seemed perfectly at ease and confident, giving her boss a wry smile as the audience settled down.

Shlomiuka studiously ignored Wilfred de Bruijn, a senior member of Jaeger's group, who sat in the audience a few rows behind Lorimer, glaring at her. On the far side of Shlomiuka sat Fenton Baumgarden, a world authority in laser physics, who would chair

the session. Baumgarden was an easy-going American who had nearly got a Nobel prize. Persistent rumours had circulated for years that it was Jaeger's reference that had spoiled it for him. But their presence together at such a prestigious event would make it clear that there was no animosity between them. Which is why Baumgarden had nominated Jaeger for the job.

Almost hidden behind the apparatus was the only other person on stage: Tony Trotman, Jaeger's head technician. It was easy to overlook Trotman — if he spoke at all it was in a near whisper, and he had a chameleon-like ability to blend into his surroundings. It would be Trotman's job to make sure the demonstration went according to plan. He waited quietly at the back of the stage, mostly staring into space.



Now and then he looked expectantly at Jaeger.

Immediately in front of the journalists were various dignitaries. Jaeger was happy to see most of them, because their influence could be useful. But two he was not pleased to see: Petra Pruszczyński and Veronique Dubois. They represented his leading competitors, and they were present only because protocol demanded it.

Pruszczyński was a whizz-kid at the Gdansk Centre for Optical Computation, and a while back had given a job to one of Jaeger's former students, the brilliant but disgraced Jirong Feng. He knew Feng would be somewhere in the room ... tucked discreetly away near the back, no doubt. Dubois was an up-and-coming experimentalist with experience in secure quantum-cryptographic communications. She worked at the University of Paris 14, in direct competition with Jaeger's group. She was friendly enough — but too cool and calculating for Jaeger to feel comfortable.

The lighting in the hall dimmed. Baumgarden, silhouetted by a single dramatic spotlight, declared the conference open and invited his colleague Dr Shlomiuka to introduce their speaker. She delivered a short, witty

introduction, entirely from memory, and sat down. Jaeger rose, to enthusiastic applause.

"Distinguished guests, colleagues, and ladies and gentlemen of the press," Jaeger began, "I am humbled and honoured to inaugurate the World Year of Physics." He paused for further applause. "I have been invited to describe my research group's new results on quantum entanglement — which Einstein so memorably described as 'spooky action at a distance'. Of course, we now understand that the phenomenon is entirely rational and not at all spooky — but it does remain quite startling. Only a few months ago my team discovered just how startling."

He glanced down at the assembled journalists. "You all know the fable of Schrödinger's cat, but it is often forgotten that Schrödinger's primary purpose was not to assert the existence of superposed quantum states, but to examine their paradoxical consequences for a macroscopic entity." He paused. "In the interests of animal welfare, I will employ a simple machine instead of a cat. I call the experiment 'Schrödinger's mousetrap'."

Dutiful laughter rippled through the audience. So that's what the model on the plinth was for. "First," Jaeger explained, "I will set up my mousetrap." He gestured towards the model. "A laser-operated optical trap. But I do not bait my mousetrap with cheese, because it has already caught a mouse. Which in this instance is a hollow dielectric sphere."

"Now, although I know there is a mouse in the trap, I do not know its quantum state." He paused. "What does that mean? It means that my spherical mouse can resonate in two distinct modes. Following the precedent set by Herr Professor Doktor Schrödinger, I shall refer to these modes as 'dead' and 'alive'."

"So have I trapped a live mouse, or a dead mouse? That, my friends, is the great mystery!"

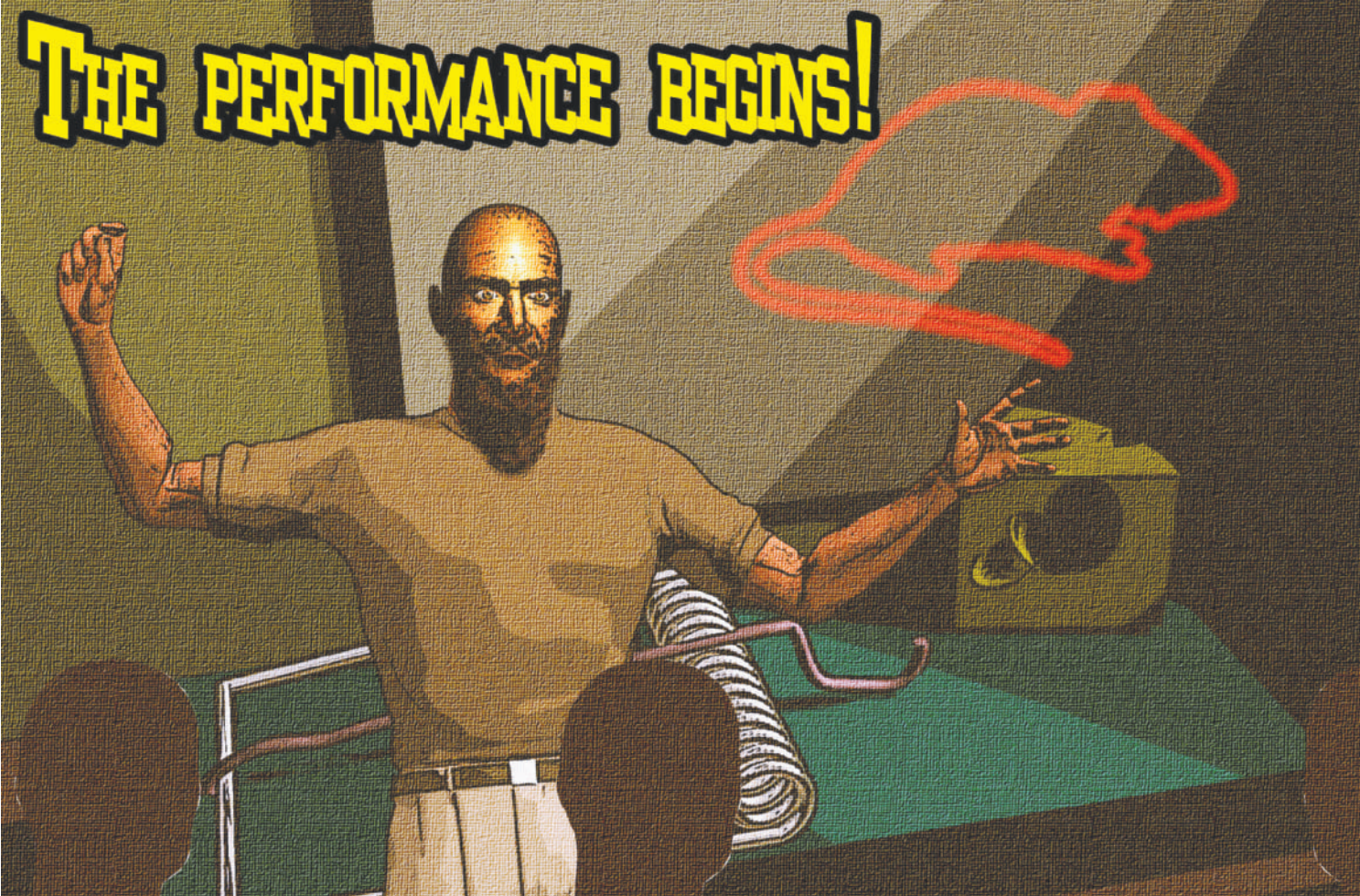
Pruszczyński turned to Dubois, and said, in a stage-whisper, "Juz' like Rufus. He iz real showman."

"Yes," Dubois whispered back. "Effective, isn't it?"

Pruszczyński glared at her, then sighed and nodded. "I am sorry to say, you are right."

"To resolve the mystery," Jaeger continued, "I shall appeal to quantum entanglement. I will arrange for the mouse's fate to be

THE PERFORMANCE BEGINS!



CHRISTIAN DARKIN

entangled with that of another sphere: the 'cat'. This sphere also resonates in two modes. The entanglement procedure will ensure that the two spheres have different states: if the cat is alive, then the mouse is dead, and vice versa. My assistant will now use a brief laser pulse to prepare the entangled states." As Trotman laboured in the darkness of the stage, the screen showed a schematic animation of the entanglement procedure.

"At the climax of the experiment," Jaeger explained, "I will observe the state of the cat, and infer the state of the mouse. If I observe the cat to be alive, then the mouse's wave function will immediately collapse to 'dead', and the mouse will stay in the trap. However, if the cat is observed to be dead, then the mouse's wave function will collapse to 'alive', and in this higher resonant mode the sphere will escape from the trap."

"How will we know if the mouse escapes or not, Professor?" one of the reporters called out.

"Ah, my impatient questioner from the press, I am coming to that. To make the results visible to everyone in this room, I have arranged the apparatus so that the laser beam will project a three-dimensional hologram of the mouse. You will see either a dead mouse caught in my trap, or a live one running across the stage above my head as it escapes."

At this point Shlomiuka noticed that Pruszcyncki was grumbling, and asked her to stand up and clarify her concerns. It would be better then letting her simmer, and eventually explode.

"Iz stupid experiment! Iz dangerous! Laser must be unnecessarily powerful, juz to make silly hologram..."

"My dear woman," said Jaeger unwisely, "there are adequate safeguards..."

"Am not your dear woman! Experiment iz totally meaningless! One trial proves nothing: iz all about correlations. Must do experiment hundreds of times!"

"Which we have done, in the laboratory, as our forthcoming paper will establish," said Jaeger smoothly. "But we don't have enough time to do that today. This is just a demonstration, Professor Pruszcyncki."

Pruszcyncki was unconvinced. "Seen preprint. Prove nothing! Extraor'nary claim require extraor'nary evidence, which you don't got!"

With ill-concealed anger, Jaeger suggested that Pruszcyncki should learn the basics of quantum entanglement. Before she could react, Baumgarten, in his role as chair, politely tried to end the discussion. But it was Shlomiuka, who always wanted to keep everyone happy, who proposed that the demonstration should be performed several times to obtain some statistical evidence. Pruszcyncki, still grumbling, sat down, while Jaeger, red-faced, dabbed at his forehead and looked a little unwell. He took his place beside the demonstration apparatus.

The lights dimmed for a ten-minute multimedia presentation of the background to the experiment and its theoretical basis. The hall and platform were now in almost total darkness, save for tiny individual read-

ing lights. The presentation ended, and Trotman readied the apparatus. The slide projector showed the cat's state: it was alive. While the audience waited for the hologram of a dead mouse to appear, a strange sizzle and a dull thud could be heard. Trotman was about to reset the apparatus for a second trial, when a terrible scream came from the stage. It was Shlomiuka.

After a few moments, someone had the presence of mind to turn the main lights on in the hall. Jaeger had collapsed, and was lying awkwardly on the stage, not moving. A faint smell of burning drifted on the air. Shlomiuka was kneeling beside Jaeger. She lifted his head and started sobbing.

Baumgarten bent down, stared at Jaeger's face, and pulled her away. "He's dead," he said blankly. "There's a hole in his head." He reached for a microphone. "Ladies and gentlemen: it seems that there has been a terrible accident."

Lorimer, who had a reputation for being calm in a crisis, rose to his feet. "Sorry to contradict you, Fenton, but has it occurred to you that what you call an 'accident' might have been deliberate?"

Baumgarten went pale. It was unlikely, but Lorimer had a point. There was only one thing to do. He instructed Security to lock the doors to the conference centre, to make sure no one could leave, and call the police.

To be continued...

Ian Stewart is at the Mathematics Institute, University of Warwick, Gibbet Hill Road, Coventry CV4 7AL, UK.

Fertility hormone in repose

James A. Dias

Egg and sperm development are triggered when follicle-stimulating hormone binds to its receptor. A three-dimensional structural snapshot reveals how the hormone slots into its receptor, and how specificity of binding is ensured.

Each year, around 20% of couples in the United States consult a fertility specialist because of problems with infertility. Many will seek help through assisted reproductive technologies¹. As part of such treatments, women may receive courses of follicle-stimulating hormone (FSH) to stimulate the recruitment and maturation of many ovarian follicles. At present, however, the hormone must be injected — a difficult and time-consuming process. Although the nature of FSH and other glycoprotein hormones has been known for more than 30 years, there is still no orally active therapeutic drug.

But such a drug might one day be developed, thanks to the findings presented on page 269 of this issue by Fan and Hendrickson². The authors have determined the three-dimensional structure of FSH bound to the large extracellular region of its receptor (FSHR_{HB}). Their work could lead to the development of small molecules with the same effects as FSH (agonists). It also opens up the possibility of developing non-steroidal FSH antagonists, for use as contraceptives in men to block sperm production, and in women to block oocyte development. On a more fundamental level, FSHR is part of the family of horseshoe-shaped leucine-rich-repeat proteins — named for a repeated motif that always contains leucine amino acids — and is the latest member of this family to have its structure determined³. It is also among the largest of the G-protein-coupled receptors (GPCRs), proteins that constitute roughly 30% of drug targets⁴, and so its structure sheds light on their mechanism of action, too.

Unlike many GPCR ligands, FSH is a highly complex molecule, and it has been studied exhaustively⁵. It is composed of two different subunits, α and β , each of which is decorated with two long chains of carbohydrate. FSH shares an identical α -subunit with three other glycoprotein hormones: luteinizing hormone, which induces testosterone production in males and the final maturation of oocytes and subsequent ovulation in females; thyroid-stimulating hormone, which regulates thyroid function; and chorionic gonadotropin, which signals pregnancy. FSH itself stimulates the initial maturation of ovarian follicles, and supports sperm production.

Given the shared α -subunit, the basis for the specificity of action of these hormones

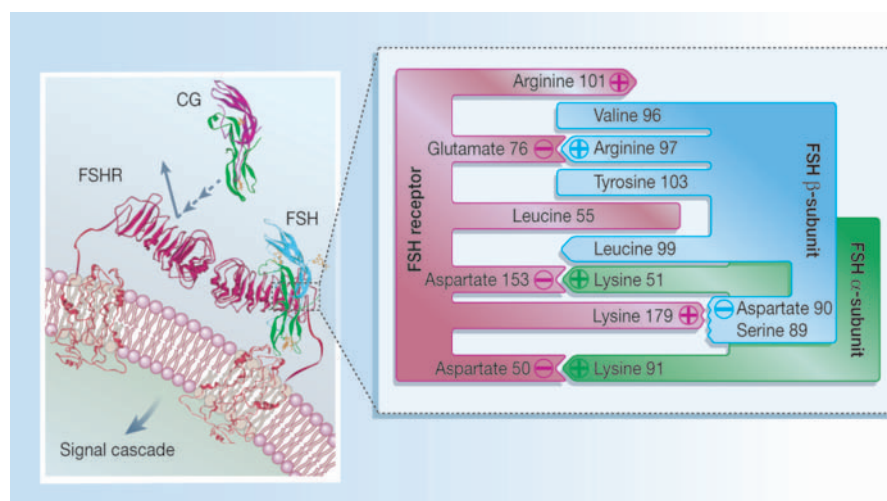


Figure 1 Hormones and receptors: keys and locks. Follicle-stimulating hormone (FSH) triggers egg and sperm production by binding to its receptor, FSHR, on the surface of granulosa cells in ovarian follicles and Sertoli cells in testes. The receptor selects FSH over related hormones (such as chorionic gonadotropin, CG, in females) to form a dimer of FSH-loaded receptors. This activates an intracellular signal cascade. Fan and Hendrickson² have determined the structure of FSH bound to the extracellular regions of FSHR; both FSH subunits contribute to the specificity of binding, with the β -subunit having the larger role. The inset shows FSH and FSHR amino acids that dictate specificity, by virtue of charge or side-chain interactions. For example, simply changing serine 89 to the positively charged arginine found in CG would destabilize the binding of FSH to FSHR.

has been a mystery, in which the dissimilar β -subunits are intuitively implicated. Without this specificity, when chorionic gonadotropin, for instance, reaches the high concentrations that it does during pregnancy, other tissues would be inappropriately — even disastrously — overstimulated. For example, ovarian hyperstimulation, a potentially life-threatening syndrome, has been associated with mutations in FSHR that bypass the specificity controls, allowing this receptor to bind to chorionic gonadotropin^{6,7}.

The mystery is now solved by the structure of the FSH-FSHR_{HB} complex². The structure shows that both hormone subunits are involved in determining the specificity of binding to the appropriate receptor, but that both charge and stereochemistry dictate specificity, with a role for α -subunit amino acids primarily in the stereochemical attributes (Fig. 1). Given the similar compositions of glycoprotein hormones, the mode of receptor binding is likely to be very similar for all of them, and so the structure of the FSH-FSHR_{HB} complex has broad impact.

Ten years ago, the structure of deglycosylated chorionic gonadotropin — the

hormone without its carbohydrate chains — was determined in two laboratories, including the Hendrickson lab^{8,9}. This began a new understanding of glycoprotein hormones, because until then it was not known how the subunits are organized into an active protein. The structure of fully glycosylated FSH¹⁰ later revealed that carbohydrate has little influence on the protein's overall structure. That structure also showed that the second loop of the β -subunit and one end of the α -subunit — its carboxy-terminal tail — are conformationally flexible.

Consistent with this previous work, Fan and Hendrickson find² that there is no carbohydrate in the binding interface of the FSH-FSHR_{HB} structure. Interestingly, a receptor-induced conformational change in FSH is also evident. FSH appears to bind to the receptor by nestling into the binding site through its β -subunit loop. Then, through a series of conformational adjustments of the neighbouring α -subunit loops, the whole molecule adopts a rigid repose that might be required for signalling.

The structure also explains the extensive data obtained by mutating FSH; these data

revealed that the α -subunit's carboxy-terminal tail is essential for receptor binding⁵. For instance, single mutations of six amino acids in this tail resulted in a 90% loss of binding activity. But it was not obvious how these amino acids contributed to binding. The structure now reveals that the α -subunit plays a neat trick: it rotates its tail by 180° so that these amino acids become fully conformationally constrained by making extensive contacts with the receptor at the hormone–receptor interface.

The structure of the FSH–FSHR_{HB} complex also provides the first structural evidence that GPCRs form dimers, which might be necessary for their full activity. In the case of FSHR, the dimer contacts are not made through the hormone. Rather, two receptors interact directly with each other, with each receptor binding one molecule of FSH. Yet the structure of the dimer presents a conundrum as well. The transmembrane regions of the two receptors in a dimer are too far apart to associate, so what drives dimer formation? One possibility is that other molecules link receptors through cytoplasmic interactions.

Of course, the next goal is to determine the structure of the full-length receptor. So far, the full-length structure of only one GPCR, rhodopsin, has been determined¹¹, and this molecule is composed solely of transmembrane domains. The role of carbohydrate in activating FSHR also needs to be defined — although receptor binding does not require carbohydrate, full activity following binding does. And it will be essential to determine the structure of FSHR without its ligand, to find out whether rearrangements occur in the receptor upon hormone binding.

Although it is in the nature of science that each advance raises new questions, the new structure² does further our understanding of the key–lock mechanism by which glycoprotein hormones work. This advance defines the specific amino acids that make contact between hormone and receptor, as well as those that repulse hormone binding and thereby provide specificity (Fig. 1). Moreover, the orientation of FSH with respect to the main extracellular regions of the receptor is now resolved, and the structure allows a guess as to how the FSH–FSHR_{HB} complex is oriented relative to the extracellular loops of the receptor's transmembrane domains — goals of researchers in the field for at least two decades.

James A. Dias is at the Wadsworth Center, New York State Department of Health, and in the Department of Biomedical Sciences, State University of New York at Albany, David Axelrod Institute, 120 New Scotland Avenue, Albany, New York 12208, USA.
e-mail: james.dias@wadsworth.org

1. Wright, V. C., Schieve, L. A., Reynolds, M. A., Jeng, G. & Kissin, D. *MMWR Surveill. Summ.* **53**, 1–20 (2004).
2. Fan, Q. R. & Hendrickson, W. A. *Nature* **433**, 269–277 (2005).
3. Kobe, B. & Kajava, A. V. *Curr. Opin. Struct. Biol.* **11**, 725–732 (2001).

4. Hopkins, A. L. & Groom, C. R. *Nature Rev. Drug Discov.* **1**, 727–730 (2002).
5. Dias, J. A. et al. in *Vitamins and Hormones* (ed. Litwack, G.) 249–322 (Academic, New York, 2002).
6. Smits, G. et al. *N. Engl. J. Med.* **349**, 760–766 (2003).
7. Vasseur, C. et al. *N. Engl. J. Med.* **349**, 753–759 (2003).
8. Laphorn, A. J. et al. *Nature* **369**, 455–461 (1994).
9. Wu, H., Lustbader, J. W., Liu, Y., Canfield, R. E. & Hendrickson, W. A. *Structure* **2**, 545–558 (1994).
10. Fox, K. M., Dias, J. A. & Van Roey, P. *Mol. Endocrinol.* **15**, 378–389 (2001).
11. Palczewski, K. et al. *Science* **289**, 739–745 (2000).

Climatology

Will soil amplify climate change?

David Powlson

It had been thought by some that rising atmospheric temperatures would have no effect on the rate at which carbon is released from the soil. A study that revisits the data behind this theory now finds otherwise.

Knorr et al.¹ in this issue (page 298) claim that rising temperatures brought about by climate change will cause microorganisms in the world's soils to decompose organic matter more rapidly, releasing extra carbon dioxide (CO₂) and accelerating climate change. This may seem unsurprising — a basic property of most biological processes is that they proceed faster with rising temperature, provided that other factors do not become limiting. However, this basic biological tenet was challenged in 2000 by Giardina and Ryan², who suggested that organic carbon decomposition in soil is not sensitive to temperature. If correct, that would mean that our understanding of the process is in serious error and predictions from current models of soil carbon turnover cannot be trusted. But Knorr et al. have re-examined the data used in Giardina and Ryan's work, and come to the opposite conclusion. The new analysis resolves the seeming paradox and suggests that the positive feedback from soil to climate might be even greater than is currently thought.

The world's soils hold about $1,500 \times 10^9$ tonnes (1,500 Gt) of organic carbon; the atmosphere contains about half this amount as CO₂ (720 Gt), and there is a further 600 Gt in vegetation³. Thus, relatively small changes in the flow of carbon into or out of soils could be significant on a global scale, and various studies have attempted to quantify these flows using models of soil carbon turnover^{4–6}.

Increasing CO₂ in the atmosphere can enhance plant growth⁷ through 'CO₂ fertilization', thus removing some of the excess CO₂. For long-lived plants such as trees, part of this carbon is sequestered for decades or centuries. With all plants, some organic carbon is transferred by roots and litter-fall into soil organic matter, some fractions of which are so strongly stabilized that they turn over on timescales of centuries or even millennia⁸. This carbon sequestration would tend to slow climate change⁹, so management practices such as re-forestation are considered as mitigation measures under the

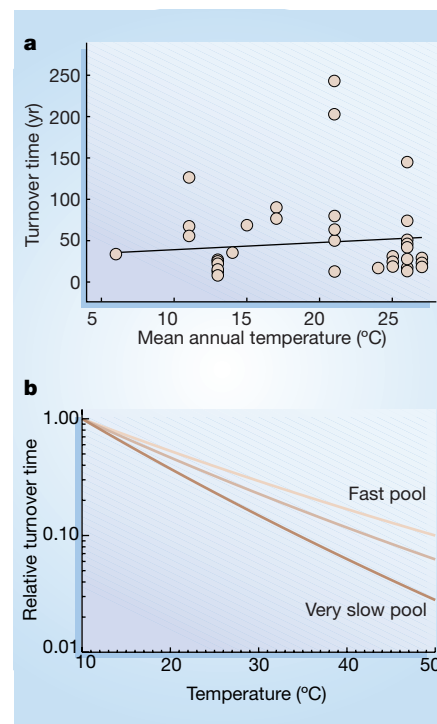


Figure 1 One question but two answers. Is the rate of decomposition of organic carbon in soil influenced by temperature? a, Giardina and Ryan² used a model with a single carbon pool, and calculated soil organic carbon turnover times that were independent of temperature. b, Knorr et al.¹ used a model containing three carbon pools (fast, intermediate and very slow). Turnover time for each pool at different temperatures is shown relative to that at 10 °C. The turnover times all decrease as temperature increases; that is, decomposition speeds up as temperature rises.

Kyoto Protocol on climate change. However, current soil models predict that, in the longer term, rising temperatures will speed up the decomposition of organic carbon in soil, releasing CO₂ into the atmosphere in excess of any carbon sequestered in the soil, and adding to climate change.

Giardina and Ryan² challenged this view with an analysis of published data from 82 experiments across five continents. They

used field measurements and laboratory incubations to calculate the turnover time of soil organic carbon. Turnover time is the inverse of the first-order rate constant for the decomposition process. They calculated that turnover times were almost independent of mean annual temperature over the range 5–35 °C (Fig. 1a).

Crucially, their analysis used the simplifying assumption that all soil organic carbon behaves in the same way and can be regarded as a single pool. Most current models, however, divide soil organic carbon into at least three pools that differ in their turnover time, and these models are fairly successful in simulating long-term changes in the organic carbon in soil¹⁰. The dangers of treating soil organic carbon as a single pool were highlighted by Davidson *et al.*¹¹, who showed that the activity of small, very active pools can mask the effects of a large, slow pool, especially in short-term experiments.

Knorr *et al.*¹ have now re-analysed Giardina and Ryan's results using a soil model with more than one carbon pool. To create their model, they used published data in which soil from a tropical rainforest in Brazil was incubated for 24 weeks at temperatures between 15 and 45 °C, and the CO₂ release was measured. They then tested how well the results were predicted by simple models of decomposition using different numbers of carbon pools, each with a temperature-sensitive rate constant. To explain the results, they found it necessary to have two active pools (one with a much faster turnover time than the other), and a very slow pool that was effectively inert during the 24 weeks of the experiment. The very slow pool was large, comprising about 95% of total soil organic carbon.

By applying this model to the incubation experiments considered by Giardina and Ryan², Knorr *et al.* demonstrate a fascinating paradox that is particularly marked if a single-pool model is assumed: because of the dominance of the large, very slow pool, short-term incubations of one to two years can seem to show no sensitivity of decomposition to temperature, even when each pool does in fact have a built-in temperature dependence. This is because, as the temperature rises, fast pools turn over even faster and their effect on the overall result is very short-lived.

By applying their model to other data sets, Knorr *et al.* reveal the perils of drawing conclusions about the impacts of climate change from short-term studies such as soil-warming experiments. Such experiments rarely continue for more than a few years, and so never provide information on the response of the large, slower pools that will dominate feedbacks from soil to atmosphere over timescales of decades or more.

Several experimental studies^{12,13} have seemed to show that decomposition is

insensitive to temperature. The reasons for this are unclear, but as Davidson *et al.*¹¹ point out, the effects of climate change are much more complex than just an increase in temperature: changes in soil water and nutrient availability will occur, and these can influence decomposition. The observation of apparent insensitivity of soil decomposition to temperature is sometimes termed acclimatization to suggest a gradual biological adaptation of soil organisms to a higher temperature. Knorr and colleagues' analysis¹ shows that biological adaptation does not have to be invoked — such results can simply be a consequence of using a model that ignores the dominance of large, slowly changing pools of organic matter.

As a final twist, Knorr *et al.* predict that, over a timescale of decades to centuries, the dominant slow pools will be more sensitive to temperature than the faster pools (Fig. 1b), causing a larger positive feedback in response to global warming than previously thought. The feedback suggested by current models is already very significant; for instance, one forecast⁶ is that organic carbon will accumulate in soil as a result of CO₂ fertilization until the middle of the twenty-first century and then decline as the effect of rising temperature

on decomposition becomes dominant.

Although Giardina and Ryan's analysis² now appears incorrect, it was useful in provoking a deeper examination of the temperature sensitivity of the decomposition of soil organic carbon^{1,11,14}, and has led to a greater appreciation of the importance of stable organic carbon fractions in influencing climate change. ■

David Powlson is in the Agriculture and Environment Division, Rothamsted Research, Harpenden, Hertfordshire AL5 2JQ, UK.
e-mail: david.powlson@bbsrc.ac.uk

1. Knorr, W., Prentice, I. C., House, J. I. & Holland, E. A. *Nature* **433**, 298–301 (2005).
2. Giardina, C. P. & Ryan, M. G. *Nature* **404**, 858–861 (2000).
3. Batjes, N. H. *Eur. J. Soil Sci.* **47**, 151–163 (1996).
4. Jenkinson, D. S., Adams, D. E. & Wild, A. *Nature* **351**, 304–306 (1991).
5. Cox, P. M., Betts, R. A., Jones, C. D., Spall, S. A. & Totterdell, I. J. *Nature* **408**, 184–187 (2000).
6. Jones, C. *et al. Glob. Change Biol.* **11**, 154–166 (2005).
7. Grace, J. & Rayment, M. *Nature* **404**, 819–820 (2000).
8. Jenkinson, D. S. *Phil. Trans. R. Soc. Lond. B* **329**, 361–369 (1990).
9. *The Role of Land Carbon Sinks in Mitigating Global Climate Change* Policy doc. 10/01 (Royal Society, London, 2001).
10. Smith, P. *et al. Geoderma* **81**, 153–225 (1997).
11. Davidson, E. A., Trumbore, S. E. & Amundson, R. *Nature* **408**, 789–790 (2000).
12. Oechel, W. C. *et al. Nature* **406**, 978–981 (2000).
13. Luo, Y., Wan, S., Hui, D. & Wallace, L. L. *Nature* **413**, 622–625 (2001).
14. Ågren, G. I. & Bosatta, E. *Soil Biol. Biochem.* **34**, 129–132 (2002).

Evolution

A taste for mimicry

Graeme D. Ruxton and Michael P. Speed

Looking inedible is a great way to deter predators, but the warning signs must be learnt first. It seems that unpalatable species employ some unexpected strategies to make the education a quick one.

Darwin saw mimicry — strong visual resemblances between unrelated species — as an excellent test case for his theories of natural selection¹. The phenomenon continues to exercise evolutionary biologists today, with the latest salvo coming from Skelhorn and Rowe². Writing in *Proceedings of the Royal Society*, they find that mimicry can work in an unexpected way to provide safeguards against predators.

Mimicry generally occurs in two forms, Batesian and Müllerian. Batesian mimicry is essentially parasitic: a prey species evolves to look like a species that is unattractive to predators (because it is poisonous, for example), and in so doing degrades the effectiveness of the signals used by the inedible species to warn off predators. By contrast, Müllerian mimicry involves two unpalatable species, and is thought to be mutualistic because the two species share the mortality costs incurred when naive predators sample them before learning to avoid the warning signal they both use³. Butterflies are among the commonest examples of Müllerian mimicry; a possible example is shown in Figure 1 (overleaf).

Müllerian mimicry has become highly contentious, with a body of mainly theoretical work arguing that systems identified previously as Müllerian might in fact be parasitic rather than mutualistic^{4,5}. But this conclusion will need to be re-examined in light of the ingenious experiments reported by Skelhorn and Rowe². An (often implicit) assumption of previous work on Müllerian mimicry has been that the two species are unappealing to predators because they have the same defence — the same toxin, say. However, there is no logical or observational foundation for this supposition⁶. Skelhorn and Rowe now demonstrate that Müllerian mimicry can provide highly effective protection from predation when the two species concerned have different defences.

The authors' experiments involved giving domestic chicks coloured food crumbs flavoured with aversive chemicals. All the crumbs were coloured identically, but some were flavoured with quinine, some with an equally bitter solution of Bitrex (a preparation designed to discourage people from biting their nails) and some with a blend of



Figure 1 Müllerian mimicry. The viceroy (*Limenitis* sp.; right) and monarch (*Danaus* sp.) butterflies look very similar but may differ in their palatability to avian predators.

the two. All flavourings produced a learnt aversion after repeated exposure. Controlling for the density of crumbs, Skelhorn and Rowe found that chicks learnt to avoid coloured crumbs more quickly when given heterogeneous populations of both quinine crumbs and Bitrex crumbs than when given homogeneous populations of all quinine, all Bitrex or all blended crumbs. Furthermore, and of equal interest, chicks retained the learnt aversion for longer when trained with the heterogeneous population than with any homogeneous population.

The authors suggest several explanations for this synergistic effect. A plausible one is that the novelty of the second taste might cause the birds to pay more attention to the associated visual signal and/or allow them to judge better how many unpalatable crumbs they have eaten. Alternatively or additionally, the distinct tastes might alert the predator to the fact that even if prey look alike, their defences are unpredictable, and the predator might therefore become more cautious. If this is true, mimicry functions to reduce a predator's attacks by providing uncertainty about the toxins in prey^{5,7}, rather than by merely sharing out the costs of initial education, as Müller originally argued. Skelhorn and Rowe's results are consistent with observations from invertebrate predators, which become more wary after they encounter prey that have toxins they've not tasted before^{8,9}. The results could also help to explain why toxins are so diverse within and among species of prey.

Regardless of the underlying mechanism, the work is a tremendous boost for the mutualistic account of mimicry. It suggests that individuals of two species with different defences can gain more protection from mimicry than from simply increasing the population density of their own species.

We hope that these results will trigger other empirical studies. In particular: do similar effects occur when there is a disparity in the levels of defence in two prey species with different defences, or even in a single prey population with only one form of defence? If heterogeneity in the amount of the same toxin produces synergistic effects in the protection from predators, this could help to explain the maintenance of

conspicuous warning displays in species that are highly heterogeneous in their levels of chemical defence.

Also, is mimicry really necessary for synergy? Repeating Skelhorn and Rowe's experiments with two types of prey that have two distinct visual signals as well as different tastes would reveal whether both visual signals are learnt more quickly when only one is encountered. If so, this would prompt a timely reconsideration of the neglected concept of a Müllerian mutualism where two defended

prey species benefit from each other even in the absence of a shared warning signal^{6,10}. Scientific interest in mimicry may be nearly 150 years old, but these fascinating experiments show that this topic still has much to tell us about evolution in the natural world. ■

Graeme D. Ruxton is at the Institute of Biomedical and Life Sciences, Graham Kerr Building, University of Glasgow, Glasgow G12 8QQ, UK.

e-mail: g.ruxton@bio.gla.ac.uk

Michael P. Speed is in the School of Biological Sciences, The University of Liverpool, Crown Street, Liverpool L69 7ZB, UK.

1. Darwin, C. *The Life and Letters of Charles Darwin* (ed. Darwin, F.) (Murray, London, 1887).
2. Skelhorn, J. & Rowe, C. *Proc. R. Soc. Lond. B* doi:10.1098/rspb.2004.2953 (2005).
3. Müller, F. *Zool. Anz.* **1**, 54–55 (1878).
4. Speed, M. P. *Anim. Behav.* **46**, 1246–1248 (1993).
5. Kokko, H., Mappes, J. & Lindström, L. *Ecol. Lett.* **6**, 1068–1076 (2003).
6. Turner, J. R. G. & Speed, M. P. *Evol. Ecol.* **13**, 807–827 (2001).
7. Sherratt, T. N., Speed, M. P. & Ruxton, G. D. *J. Theor. Biol.* **228**, 217–226 (2004).
8. Pasteels, J. M. & Gregoire, J. C. *J. Chem. Ecol.* **10**, 1693–1700 (1984).
9. Rayor, L. S. & Munson, S. *Entomol. Exp. Appl.* **104**, 193–201 (2002).
10. Leimar, O. & Tuomi, J. *J. Evol. Ecol.* **12**, 59–71 (1998).

Astronomy

Weighing the baby

I. Neill Reid

Mass is the fundamental parameter in stellar astrophysics, but measuring mass is difficult, especially for young stars. A study of a youthful neighbour of the Sun provides insight into the accuracy of widely used calibrations.

Size matters for stars. Individually, a star's mass determines which fusion reactions are possible in the core, and hence its luminosity, surface temperature and the length of its life. For example, objects less massive than approximately 0.075 times the mass of the Sun (about 75 Jupiter masses) never ignite hydrogen fusion, and therefore fade to obscurity in 100 million years or less; these are brown dwarfs. Collectively, the 'initial mass function', the number of stars and brown dwarfs forming as a function of mass, is critical in assessing the total amount of matter in the Milky Way. Until recently, this task loomed large, but within the past five years it has become clear that brown dwarfs are not viable candidates for 'dark matter', the invisible material that constitutes some 90% of the mass of the Milky Way. The initial mass function is also important for testing theories of star formation, and remains vital for understanding how the material in molecular clouds is redistributed to form stars. In particular, we would like to know whether there is a low-mass cut-off to the formation process.

Unfortunately, we cannot measure stellar mass directly, and although there are indirect techniques for accurately determining the

mass of mature stars, these are unreliable for young stars. Elsewhere in this issue (*Nature* **433**, 286–289; 2005), Laird M. Close and his collaborators report the first precise determination of the mass of a small star. Their findings may have important implications for our understanding of the low-mass cut-off for star formation and the estimated number of brown dwarfs in young clusters such as the Orion nebula.

To weigh a star, we have to determine the effects of its mass. Thus, we measure how mass regulates the orbits of objects in binary-star or planetary systems; how it affects the wavelength of light through gravitational redshift (about 0.5 km s⁻¹ for the Sun); and its effects on the brightness, shape and/or location of distant objects in gravitational lensing. Each of these techniques can be applied only to a small number of objects, but luminosity and mass are correlated in mature hydrogen-burning stars, which follow a well-defined evolutionary sequence; as a result, we can map out a mass–luminosity relation using the relatively small number of stars in binary systems whose masses have been accurately determined. It is this relationship that provides the basis for the derivation of the mass function of stars in the solar neighbourhood.

The mass–luminosity calibration, however, is not appropriate either for brown dwarfs, which never reach a stable configuration, or for young stars (those less than 100 million years old), which are still contracting onto the well-defined sequence followed by mature stars and therefore have larger radii and higher luminosities. This is unfortunate, because clusters of young stars offer the best hunting ground for low-mass stars and brown dwarfs exactly because those objects are more luminous and therefore easier to identify and characterize. Star-forming regions such as the Orion nebula (which is about 5 million years old) provide the best opportunity for identifying isolated brown dwarfs that have planet-size masses. However, the masses of individual objects can be derived only using theoretical models, matching the observed luminosities and temperatures against predicted mass–luminosity relations for the appropriate age. These analyses generally indicate that the mass function is turning over in the brown-dwarf regime, with perhaps 30% as many brown dwarfs as stars; moreover, a few studies have claimed to detect objects with masses as low as 1–2 Jupiter masses, suggesting that we have yet to reach a lower mass limit.

Close and his collaborators have provided the first empirical test of theoretical mass–luminosity relations at low masses. The object of their studies is AB Doradus, which lies only 50 light years away and is about 50 million years old — and hence a celestial infant compared with the 5-billion-year-old Sun. AB Doradus proves to be a quadruple star system. It is composed of two close pairs, AB Dor A and C, and AB Dor Ba and Bb, each separated by 1 to 2 astronomical units (AU; 1 AU is the distance between the Earth and the Sun). The two pairs lie 135 AU apart.

Close *et al.* have used high-resolution imaging to map the orbit of the AB Dor A and C system. Although this is not the first time that dynamical masses have been derived for young objects, all previous investigations have centred on stars with masses comparable to that of the Sun. In contrast, AB Dor C has a mass of 0.09 solar masses — only slightly above the brown-dwarf limit. Crucially, matching AB Dor C's observed properties against theoretical models shows that the 0.09-solar-mass model systematically overestimates both the luminosity and the temperature. In other words, had we used the theoretical models to analyse the observed colours and magnitudes, we would have significantly underestimated the mass and concluded that AB Dor C was a brown dwarf.

Clearly, this is only one datum point, for one particular age and one particular mass. Nonetheless, there are wide potential ramifications. If this is a fair reflection of the theoretical models, then analyses of the luminosity distribution of low-mass stars and brown dwarfs in young clusters may be

systematically underestimating their masses. In consequence, the turnover in the mass function would be exaggerated (that is, there would be more brown dwarfs than currently estimated), but a lower-mass cut-off might lie at higher masses, because the supposed objects of 1–2 Jupiter masses might in fact exceed 15–20 Jupiter masses (0.015–0.02 solar masses).

Systems such as AB Doradus are few and

far between. But Close and colleagues' result emphasizes the importance of using them to add more empirical reference points, and to thoroughly check the accuracy of theoretical models for determining stellar mass. ■

I. Neill Reid is at the Space Telescope Science Institute, Science Policies Division, 3700 San Martin Drive, Baltimore, Maryland 21218, USA.
e-mail: inr@stsci.edu

Signal transduction

A new canon

Olivier Pourqu  

Muscle development in vertebrates relies on signals transmitted from proteins of the Wnt family. But which molecules form the relay that transfers this signal to the cell nucleus? The answer is unexpected.

The discovery of interconnections between signalling pathways that were once thought to be distinct is becoming a familiar theme in biology. Even so, aficionados may be surprised at the link, described by Chen *et al.*¹ on page 317 of this issue, between some of the most familiar proteins involved in development, and one of the best-characterized responses to hormones and other extracellular signals.

Among the cellular transmembrane receptors that are involved in transmitting external stimuli to the cell nucleus, the family of G-protein-coupled receptors (GPCRs) is by far the largest and most diverse in animals, numbering hundreds of members in humans². These receptors can recognize stimuli ranging from peptide hormones to odorant molecules, and are characterized by a structure that comprises seven transmembrane regions.

The classical intracellular pathway that is induced by the GPCRs starts with the activation of a trimeric G protein associated with the receptor. Specifically, the guanine triphosphate (GTP) molecule that is bound to the G protein's α -subunit is hydrolysed; the α -subunit then becomes detached and regulates the activity of the enzyme adenylyl cyclase. Activated adenylyl cyclase converts adenosine triphosphate (ATP) into cyclic adenosine monophosphate (cAMP), which in turn binds and activates an enzyme called protein kinase A (PKA). The catalytic subunit of PKA enters the nucleus and phosphorylates the gene-transcription factor CREB, which activates genes that contain cAMP-response elements³.

The typical Wnt signalling cascade — a major pathway in development and disease — is rather different. Proteins of the Wnt family are secreted molecules that bind to a particular class of seven-transmembrane-domain receptors, the Frizzled proteins, which are related to GPCRs⁴. In the best-

characterized intracellular pathway that occurs following Wnt binding — the 'canonical' pathway — the β -catenin protein becomes stabilized through inhibition of the Gsk3 protein kinase. This allows β -catenin to enter the nucleus, where it cooperates with transcription factors of the Lef/TCF family. Non-canonical Wnt pathways that are involved in various embryonic patterning processes (the control of planar cell polarity, for instance) have been suggested to involve G proteins such as Rho. But the classical cAMP pathway that lies downstream of G proteins had never been seen to be activated by Wnt binding to their receptors — until now.

In the embryo, the development of vertebrate skeletal muscles, a process known as myogenesis, begins in transient blocks of tissue called somites. During the first steps of myogenesis, the dorsal region of the somites forms the dermomyotome, a structure that will yield all of the body's skeletal muscles⁵. These early stages of muscle differentiation have been shown in chick and mouse embryos to require several secreted molecular signals, such as Wnt1, produced by the dorsal neural tube, and Wnt7A, produced by the surface ectoderm tissue⁶. But the intracellular pathway triggered by these signals remained elusive. Now Chen *et al.*¹ report that the classical cAMP pathway is involved.

Using immunohistochemistry, the authors initially observed that the phosphorylated form of CREB is enriched at sites of ongoing myogenesis such as the dermomyotome. This prompted them to look at muscle differentiation in the absence of CREB. This they did using CREB-deficient mice, as well as mouse embryos infected with a virus encoding a mutant form of CREB — or of PKA — that impeded the function of the normal protein. In all of these cases, certain proteins (Pax3, MyoD and Myf5) that serve as markers of developing muscle were no longer expressed. The differentiation of

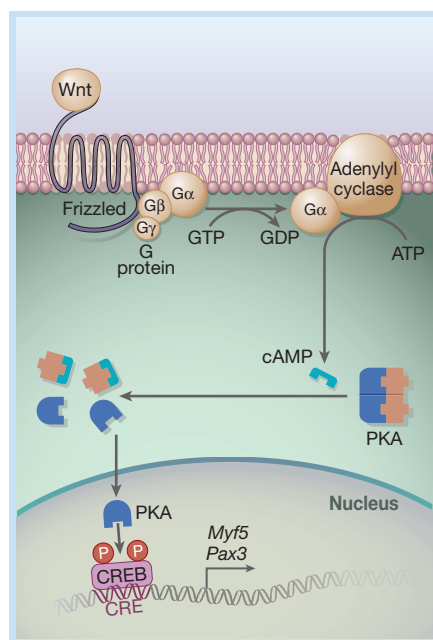


Figure 1 Breaking the mould: a new pathway downstream of the Wnt proteins. Chen *et al.*¹ find that muscle development in vertebrates involves the well-known cyclic AMP (cAMP) pathway downstream of Wnts. This proposed pathway begins when a Wnt protein activates its receptor, Frizzled, which is bound to a trimeric G protein. The G protein hydrolyses GTP to GDP, resulting in the release of the G protein's α -subunit. This $G\alpha$ subunit activates adenylyl cyclase, which in turn generates cAMP from ATP; the cAMP binds to protein kinase A (PKA), changing its conformation such that its catalytic subunits are released and move into the nucleus. There they phosphorylate the transcription factor CREB, which binds to genes containing a cAMP-response element (CRE). These genes might include *Myf5* and *Pax3*, markers of muscle development.

regions of the somite destined to form vertebrae was not affected, however.

Chen *et al.* then used a system in which myogenesis can be activated by culturing muscle precursors with cells expressing Wnt1 or Wnt7A. They found that the induction of myogenesis is accompanied by phosphorylation of CREB. Moreover, in this experimental system, inhibiting PKA or CREB blocks the onset of myogenesis. The authors also found that Wnt signals can increase the activation of genes containing a CREB-responsive element. And genes that control myogenesis, such as *Pax3* and *Myf5*, possess such elements, indicating that they might be regulated by CREB (although it remains to be seen whether they are direct targets of the pathway). Together, these results provide conclusive evidence that the classical cAMP pathway acts downstream of Wnt binding to control myogenesis (Fig. 1).

The discovery of this signalling connection has numerous implications — for

myogenesis to begin with. Shh is another major signalling molecule that controls the developmental patterning of somites⁷. During the initial stages of this process, Shh can antagonize Wnt-induced expression of *Pax3*, a dermomyotome marker. Conversely, Wnt can antagonize Shh-induced expression of *Pax1*, which marks the precursors of the vertebrae. Given that Shh signalling is inhibited by PKA, PKA activation by Wnts may provide them with a means of antagonizing Shh, to establish the dermomyotome domain precisely.

Paradoxically, however, Shh and Wnt are thought to act synergistically to allow the next tissue involved in muscle development — the myotome — to emerge from the edges of the dermomyotome. This might be explained by the finding that the high PKA activity in the dermomyotome edges falls markedly as dermomyotome cells take up residence in the myotome proper and become more distant from Wnt-producing cells. Such a stepwise coordination of Wnt and Shh signalling could form the basis of their synergism.

Given that the Frizzled proteins are seven-transmembrane-domain receptors, Chen and colleagues' findings¹ also strengthen the idea that they can act as bona fide GPCRs, signalling through G proteins to cAMP and beyond — an idea that had been in question. Much of our understanding of the pathways activated by Wnts has come from genetic analyses in fruitflies⁸, and components of the cAMP pathway have not been picked up in exhaustive screens of signalling that target a key member of the fruitfly Wnt family called Wingless. Moreover, in fruitflies, the cAMP pathway has been implicated in the control of several behavioural traits⁹, but had not been associated with Wnt-regulated developmental processes. That, too, is changing, however: Katanaev *et al.*¹⁰ have just reported a role for G proteins in Wnt signalling in fruitflies. So it is plausible that Chen and colleagues' findings extend far beyond muscle development, and that a completely new chapter on Wnt signalling needs to be written.

Olivier Pourquié is at the Stowers Institute for Medical Research, 1000 E 50th Street, Kansas City, Missouri 64110, USA.

e-mail: olp@stowers-institute.org

1. Chen, A. E., Ginty, D. D. & Fan, C.-M. *Nature* **433**, 317–322 (2005).
2. Vassilatis, D. K. *et al. Proc. Natl Acad. Sci. USA* **100**, 4903–4908 (2003).
3. Shaywitz, A. J. & Greenberg, M. E. *Annu. Rev. Biochem.* **68**, 821–861 (1999).
4. Barnes, M. R., Duckworth, D. M. & Beeley, L. J. *Trends Pharmacol. Sci.* **19**, 399–400 (1998).
5. Gros, J., Scaal, M. & Marcelle, C. *Dev. Cell* **6**, 875–882 (2004).
6. Tajbakhsh, S. *et al. Development* **125**, 4155–4162 (1998).
7. Fan, C. M. *et al. Cell* **81**, 457–465 (1995).
8. Moon, R. T., Bowerman, B., Boutros, M. & Perrimon, N. *Science* **296**, 1644–1666 (2002).
9. Hendricks, J. C. *et al. Nature Neurosci.* **4**, 1108–1115 (2001).
10. Katanaev, V. L., Ponzelli, R., Sémériva, M. & Tomlinson, A. *Cell* **120**, 111–122 (2005).



100 YEARS AGO

India. By Colonel Sir Thomas Holdich.

With climates varying from the ice-bound deserts of the higher Himalayas and the rain-steeped forests of Tenasserim, to the desolation of Makran... where in one part music is produced by stamping on a piece of wood, and in another has been carried to a refinement which requires sixty-four tones to our octave... where there is found a system of laws so elaborate that the cashier who has confessed to embezzlement may yet succeed in escaping punishment, and a system of government so paternal that it imprisons the husband, whose domestic happiness has been ruined, to prevent his committing the crime of murder; the territories known as British India may be a country for political purposes, but in no proper sense of the word do they constitute a nation... To write a description which, in a book of moderate compass, will convey a clear and fairly proportioned conception, requires a master hand; not to have failed is in itself high praise, but Sir Thomas Holdich has done more than this, he has produced a topographical description of the Indian Empire which... is not only interesting to read, but accurate and well proportioned on the whole. From *Nature* 19 January 1905.

50 YEARS AGO

In a publication entitled "Jean-Sylvain Bailly, Astronomer, Mystic, Revolutionary, 1736–1793"... Edwin Burrows Smith gives an account, including a very comprehensive bibliography and index, of one who was a statesman, astronomer and savant... Born into an essentially artistic milieu, Bailly abandoned the arts to study Newtonian physics, became a respected astronomer and won wide recognition throughout Europe for his research on Jupiter's satellites... The study of Newtonian physics led Bailly to the belief that scientific precision could be extended to other fields of human knowledge, such as history, law, language, etc., and this view led to a breach with the sceptical philosophers who had little confidence in systematic explanations of immaterial or intangible phenomena. As the author of this biography remarks, "Bailly fell victim to the hobgoblin of *vraisemblance*"; in other words, in cases where truth could not be demonstrated, he was prepared to accept the probable explanation. The French Revolution represented the acid test for Bailly's philosophy... and Bailly's ideas did not survive the test. From *Nature* 22 January 1955.

Self-assembly

Algorithmic crystals made to order

PLoS Biol. **2**, 2041–2053 (2004)

Erwin Schrödinger famously predicted that biological information is encoded in an 'aperiodic crystal', which turned out to be DNA. Paul W. K. Rothmund *et al.* now show that DNA's 'digital logic' can be used to create an aperiodic crystal that grows in two dimensions in the form of a fractal pattern called a Sierpinski triangle. This object is made up of an orderly array of triangles of all sizes, and can be approximated by building the triangles from discrete 'tiles'. The triangular pattern can be created from tiles of two colours ('black' and 'white') by imposing local rules that govern the colours of neighbouring tiles.

Rothmund *et al.* embody these rules in artificial DNA molecules with dangling single strands that can pair with complementary strands on other tiles. They are either 'white' or 'black' depending on whether or not they contain loops that show up as bright regions when the resulting assemblies are scanned with an atomic force microscope. The DNA tiles do indeed form structures with the Sierpinski-triangle form, albeit fragmented by occasional tiling mismatches which, once formed, introduce erroneous information that propagates from one row to the next.

These molecules thus embody, at the molecular scale, the theoretical entities called cellular automata, which can form complex patterns depending on their pairing rules. In effect, the large-scale, hierarchical structure of the assemblies is determined by a computational algorithm programmed into the building blocks.

Philip Ball

Particle physics

They do it with lasers

Phys. Rev. Lett. **93**, 263401 (2004)

The antihydrogen quest continues.

C. H. Storry *et al.* present a laser-controlled process for creating these anti-atoms, each of which comprises an antiproton and a positron, and is hence the antimatter mirror of hydrogen.

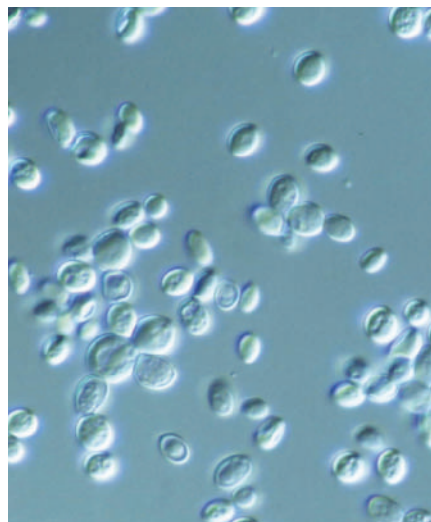
The first anti-atoms — nine in total — were created at CERN, Geneva, in 1995; by 2002, a more significant quantity had been generated. The long-term aim is to make a precise spectroscopic comparison of antihydrogen and hydrogen, to investigate the asymmetry between matter and antimatter.

Storry *et al.* — the ATRAP collaboration — demonstrate a way of making antihydrogen that has the advantage of controlling the energy of the state in which

the anti-atoms are created. They use lasers to excite caesium atoms, which, in collision with a positron, lose an electron to produce an excited state of positronium (a bound electron–positron pair). The positronium then collides with an antiproton; the electron is released and the antiproton captures the positron, forming antihydrogen.

Importantly, the anti-atom is in the same laser-selected excitation state as the original caesium atom and probably inherits the low velocity of the antiproton from which it formed. The next step is to show that de-excitation can produce cold, trapped anti-atoms for spectroscopic tests.

Alison Wright



Molecular evolution

Lab yeast on the fast track

Proc. Natl Acad. Sci. USA

doi:10.1073/pnas.0409159102 (2005)

Around 70 years ago, a strain of the yeast *Saccharomyces cerevisiae* was isolated from a rotten fig — and has been used in experiments ever since. Now, Zhenglong Gu *et al.* show that this lab strain has a faster rate of evolution than its wild counterpart.

The authors compared the entire genome of a wild yeast strain isolated from an AIDS patient's lungs with those of lab *S. cerevisiae* (pictured) and a closely related species, *S. paradoxus*. They found that the lab strain has accumulated many more genetic mutations than the wild strain, especially of a type that alters the amino-acid sequence of encoded proteins. These mutations crop up throughout all genes, suggesting that they are not a result of increased selective pressure on a particular group of genes associated with lab survival.

The findings show that the laboratory strain is evolving faster than its wild counterpart, the authors say, probably because selective pressures have relaxed during lab life. The discovery reinforces the idea that laboratory strains do not precisely

mirror the biology of their wild cousins, something that should be borne in mind when examining gene function.

Helen Pearson

Materials chemistry

Shaping silica softly

J. Am. Chem. Soc. **127**, 325–330 (2005)

Nature is renowned for its 'soft processing' of materials. Taking a leaf from nature's book, Kristian M. Roth *et al.* have identified a small molecule capable of precipitating silica from soluble silicon-containing precursor compounds.

A similar process occurs in the marine sponge *Tethya aurantia*, which uses an enzyme called silicatein to form its silica spicules. From screening studies, Roth *et al.* found a small molecule, cysteamine, which contains two key chemical groups like those in the active site of silicatein. Using this catalyst, they could encapsulate fluorescent proteins, and even living bacteria engineered to express such proteins, in a silica matrix. These encapsulated biomaterials could be useful for sensing and catalytic applications.

Conventional (sol–gel) methods for making silica structures from solution use harsh reagents. But the biomimetic catalyst requires only the mildest of conditions.

Philip Ball

Cancer

Immune investigations

J. Exp. Med. doi:10.1084/jem.20041379; 10.1084/jem.2004137 (2005)

One treatment hope for melanoma, a particularly deadly form of cancer, is to vaccinate patients with proteins present on the tumour itself. But a closer look at how such vaccines might work suggests that the process is more complex than expected.

Pierre G. Coulie, Thierry Boon and colleagues vaccinated melanoma patients with a tumour protein called MAGE-3. The aim was to produce a large number of immune cells called T cells that could destroy the tumour. Yet even in patients whose tumours regressed, levels of anti-MAGE-3 T cells in the blood were confusingly low — less than 0.001% of T cells. The numbers of T cells that target other tumour proteins, however, were much higher, both before and after vaccination.

In a second paper, the researchers found that — unlike the anti-MAGE-3 T cells — the other anti-tumour T cells were highly enriched in the metastases of a melanoma patient. Moreover, new anti-tumour T cells appeared after vaccination. The authors suspect that the existing T cells are part of a previous spontaneous anti-tumour response that became ineffective, and that they are roused from slumber by the small, yet significant, numbers of T cells raised directly against the vaccine.

Michael Hopkin

Choosing when to be a cleaner-fish mimic

A dangerous fish can discard a seemingly harmless disguise to suit its circumstances.

Mimicry in vertebrates is usually a permanent state — mimics resemble and normally accompany their model throughout the life stages during which they act as mimics. Here we show that the bluestriped fangblenny fish (*Plagiotremus rhinorhynchus*), which aggressively attacks other coral-reef fish, can turn off the mimetic colours that disguise it as the benign bluestreak cleaner wrasse, *Labroides dimidiatus*, and assume a radically different appearance. This opportunistic facultative mimicry extends the fangblenny's scope by allowing it to blend into shoals of small reef fish as well as to remain inconspicuous at cleaning stations.

The bluestreak cleaner wrasse is a cleaner fish that removes ectoparasites from larger fish clients¹. Juveniles sport a single electric-blue stripe along each side of a black body. A fangblenny in mimicking mode is identical in coloration to juvenile cleaners (Fig. 1a, b) and it uses its deceptively benign appearance to ambush other fish, lunging at them and tearing away their tissue and scales with its large canines^{2,3}.

Mimetic coloration enables fangblennies that associate with juvenile cleaners to encounter and attack more victims, but does not confer advantages in the absence of cleaners⁴. Fangblennies do occur in non-mimetic patterns, namely with a brown-to-olive or bright orange body and a pair of blue lateral stripes^{5,6} (Fig. 1c, d); these non-mimetic fangblennies also ambush fish⁶.

We examined the context of this colour variation in fangblennies on Indonesian reefs

(for methods, see supplementary information), predicting that non-mimetic fangblennies would be less likely to spend time alongside a cleaner fish than would mimetic ones. We also investigated whether the polymorphism is the result of fixed inter-individual differences or of individual plasticity.

The adoption of mimetic or non-mimetic colour patterns seems to be linked to social context. Mimetic fangblennies were significantly more likely to be seen with juvenile cleaner wrasse than were non-mimetic fangblennies ($\chi^2_1 = 11.5$, $P < 0.001$; Fig. 2). By contrast, non-mimetic fangblennies were more often found among shoaling fish than were mimetic fangblennies ($\chi^2_1 = 5.0$, $P = 0.03$; Fig. 2). The most common shoal-mates of non-mimetic fangblennies were red-cheeked and purple anthias (*Pseudanthias huchti* and *P. tuka*, respectively) and shoulder-spot wrasses (*Leptojulis cyanopleura*), which form loose shoals just above the reef and provide effective cover for non-mimetic fangblennies (at least to the human eye).

Two lines of evidence suggest that individual fangblennies can change their colour at will. First, the mean body sizes of fangblennies initially encountered in the black, olive/brown and orange patterns were similar ($F_{2,56} = 1.61$, $P = 0.22$), and the size distributions of all three colours overlapped nearly completely, suggesting that the patterns are not simply ontogenetic shifts. Second, fangblennies altered their colour quickly during translocation experiments. All four translocated mimetic fangblennies adopted an alternative colour (one orange, three

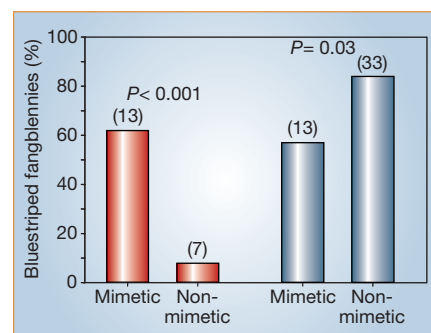


Figure 2 Percentage of bluestriped fangblennies in mimetic (black with one blue stripe) and non-mimetic (olive/brown or orange, with two blue stripes) colour patterns associated with juvenile cleaner wrasses (red bars) and with shoaling small reef fish (blue bars); there were 21 mimetic fangblennies and 38 non-mimetic fangblennies (some were associated with both juvenile cleaner wrasses and shoals, and some were associated with neither, hence totals do not sum to 100%). The number of fangblennies contributing to each category is given in parentheses. P values were derived from χ^2 tests.

olive/brown) and striping pattern (from one to two stripes on each side) within minutes of release and maintained this pattern for up to three hours.

Three of five non-mimetic fangblennies also changed colour to a different non-mimetic pattern (two orange to olive/brown, one olive/brown to orange). Although these changes may have been induced by laboratory stress, they still indicate that rapid colour change in individuals is possible and that new colours can be maintained for several hours. However, the physiological basis of these colour changes is not yet understood.

Our results indicate that bluestriped fangblennies are facultatively mimetic and that they may adjust their appearance to the presence or absence of cleaner fish. Mimetic fangblennies can operate in exposed locations near cleaning stations, whereas non-mimetic fangblennies often conceal themselves among fish shoals from which they can strike at passing fish. These tactics are equally successful: mimetics made 2.8 ± 3.1 attacks per 15 min and non-mimetics made 3.6 ± 4.4 attacks per 15 min ($t_{65} = 0.77$, $P = 0.45$). Plasticity in mimetic coloration releases mimics from dependence on the presence of a model and allows them to persist in a non-mimetic form on reefs from which cleaners have been lost or on which cleaners have never settled.

Isabelle M. Côté, Karen L. Cheney

Centre for Ecology, Evolution and Conservation,
School of Biological Sciences, University of East
Anglia, Norwich NR4 7TJ, UK

e-mail: i.cote@uea.ac.uk

1. Grutter, A. S. *Nature* **398**, 672–673 (1999).
2. Russell, B. C., Allen, G. R. & Lubbock, H. R. *J. Zool.* **180**, 407–423 (1976).
3. Kuwamura, T. *Nanki Seibutu* **23**, 61–70 (1981).



Figure 1 Facultative mimicry by bluestriped fangblennies (*Plagiotremus rhinorhynchus*; family Blenniidae). **a**, Fangblenny with mimetic colour pattern; **b**, the mimic model, a juvenile bluestreak cleaner wrasse (*Labroides dimidiatus*; family Labridae); **c**, d, fangblennies with non-mimetic coloration (**c**, orange form; **d**, olive/brown form).

4. Côté, I. M. & Cheney, K. L. *Proc. R. Soc. Lond. B* **271**, 2627–2630 (2004).
 5. Smith-Vaniz, W. F. *Acad. Nat. Sci. Philadelphia Monogr.* **19**, 1–196 (1976).
 6. Randall, J. E., Allen, G. R. & Steene, R. *Fishes of the Great Barrier Reef and Coral Sea* (Crawford House, Bathurst, Australia, 1997).
- Supplementary information accompanies this communication on Nature's website.
- Competing financial interests: declared none.

Behavioural ecology

Transient sexual mimicry leads to fertilization

Sexual mimicry among animals is widespread^{1,2}, but does it impart a fertilization advantage in the widely accepted 'sneak-guard' model³ of sperm competition? Here we describe field results in which a dramatic facultative switch in sexual phenotype by sneaker-male cuttlefish leads to immediate fertilization success, even in the presence of the consort male. These results are surprising, given the high rate at which females reject copulation attempts by males, the strong mate-guarding behaviour of consort males, and the high level of sperm competition in this complex mating system^{4,5}.

The giant Australian cuttlefish, *Sepia apama*, is a solitary cephalopod mollusc that forms large aggregations to mate and lay eggs⁴. Females lay one large egg at a time (5–39 per day) and mate up to 17 times with 2–8 males daily. The operational sex ratio averages four males to one female, but ranges up to eleven males to one female. Females reject 70% of mating attempts and the competition between males for mates is intense; mate guarding is almost continuous. Consort males obtain 64% of matings, the remainder being by small, unpaired or extrapair males. Small males (with male coloration and of similar size to females) obtain extrapair copulations by 'open' stealth (approaching a guarded female as the consort is repelling other males), by 'hidden' stealth (meeting females under rocks) or — the subject of this investigation — by mimicking the appearance and behaviour of females^{4,6}.

Visual deception is achieved when small males suddenly hide their sexually dimorphic fourth arms, acquire the mottled skin patterning typical of females, and shape their arms to mimic the posture of egg-laying females, who are not receptive to mating⁴ (Fig. 1; for video, see supplementary information). This facultative change in appearance is instantaneous and occurs at a rate of about 10 changes per 15 min during intense behavioural interactions (demonstrated in 12 video subsamples (data not shown); mimic duration, 10–184 s). We found that female mimickers could successfully deceive the consort male and that they were able to position themselves near the female in 30 out of 62 attempts. Other males attempted to

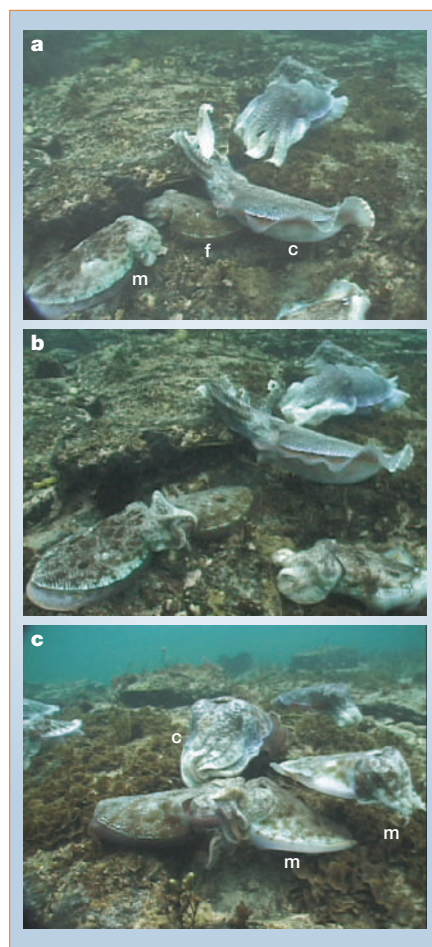


Figure 1 Female mimicry by 'sneaker' male cuttlefish leads to extrapair copulations and fertilization. **a**, An unpaired male (m) assumes female coloration and posture and approaches the paired female (f) while the consort (c) displays to an approaching large male (top right); note the conspicuous, large white arms, a sexually dimorphic male character. **b**, The female accepts a mating attempt by the female mimic as the consort continues to display to the other large male. **c**, The consort male allows the mimic to finish mating without interruption, even when he is not distracted (for video, see supplementary information); the other small 'sneaker' male has swiftly transformed into a female mimic as well.

mate with the mimics 41 times (25 attempts by large males and 16 by small males, two of which were by other female mimics).

We observed five initiations of mating by mimics. One mimic was rejected, one was interrupted by the consort male, and three resulted in successful spermatophore transfer. Consorts did not mate with females immediately after being cuckolded by the mimic. We tested the paternity of the next laid egg by using five microsatellite DNA fingerprints⁵ on tissue samples from the females and any male that mated with them during the focal samples (up to four in 30 min; females had up to 5 male genotypes in their sperm sources, which included recently attached spermatangia and stored sperm⁷). Two of the three successful inseminations by mimics resulted in fertilization. Both guarders and sneaks are able to father subsequent eggs⁵, so the matings by mimics may

have fertilized later eggs, but we could not test this possibility because SCUBA-diving constraints curtailed our sampling efforts.

Cuttlefish have keen vision but poor social recognition⁸, which favours visual sexual mimicry. In contrast to taxa in which alternative mating tactics are either fixed or vary ontogenetically^{1,9,10}, cuttlefish use neural control to change their skin patterning, posture and tactics instantly. To our knowledge, this is the first demonstration of immediate fertilization success in an animal using facultative mimicry as part of a conditional mating strategy¹¹. These field results, combined with those on bluegill sunfish¹⁰ (*Lepomis macrochirus*), which are 'broadcast' spawners, lend genetic and behavioural support to Parker's predictions that some sneaks can win sperm competition and steal³ fertilization from guarders.

Roger T. Hanlon*, Marié-José Naud†‡, Paul W. Shaw†, Jon N. Havenhand‡§

*Marine Resources Center, Marine Biological Laboratory, Woods Hole, Massachusetts 02543, USA
e-mail: rhanlon@mbi.edu

†School of Biological Sciences, Royal Holloway College, University of London, Egham TW20 0EX, UK

‡School of Biological Sciences, Flinders University, Adelaide 5001, Australia

§Tjärnö Marine Biological Laboratory, Gothenburg University, S45296 Strömstad, Sweden

1. Andersson, M. *Sexual Selection* (Princeton Univ. Press, Princeton, New Jersey, 1994).
2. Saetre, G. P. & Slagsvold, T. *Am. Nat.* **147**, 981–995 (1996).
3. Parker, G. A. in *Sperm Competition and Sexual Selection* (eds Birkhead, T. R. & Møller, A. P.) 3–54 (Academic, London, 1998).
4. Hall, K. C. & Hanlon, R. T. *Mar. Biol.* **140**, 533–545 (2002).
5. Naud, M. J., Hanlon, R. T., Hall, K. C., Shaw, P. W. & Havenhand, J. N. *Anim. Behav.* **67**, 1043–1050 (2004).
6. Norman, M. D., Finn, J. & Tregenza, T. *Proc. R. Soc. Lond. B* **266**, 1347–1349 (1999).
7. Naud, M. J., Shaw, P. W., Hanlon, R. T. & Havenhand, J. N. *Proc. R. Soc. Lond. B* (in the press).
8. Boal, J. G. *Anim. Behav.* **52**, 529–537 (1996).
9. Shuster, S. M. *Evolution* **43**, 1683–1698 (1989).
10. Fu, P., Neff, B. D. & Gross, M. R. *Proc. R. Soc. Lond. B* **268**, 1105–1112 (2001).
11. Gross, M. R. *Trends Ecol. Evol.* **11**, 92–98 (1996).

Supplementary information accompanies this communication on Nature's website.

Competing financial interests: declared none.

brief communications arising online

www.nature.com/bca

Evolutionary genomics: Codon bias and selection on single genomes

M. W. Hahn, J. G. Mezey, D. J. Begun, J. H. Gillespie, A. D. Kern, C. H. Langley, L. C. Moyle
(doi:10.1038/nature03221)

Evolutionary genomics: Detecting selection needs comparative data

R. Nielsen, M. J. Hubisz (doi:10.1038/nature03222)

Evolutionary genomics: Codon volatility does not detect selection

Y. Chen, J. J. Emerson, T. M. Martin
(doi:10.1038/nature03223)

Reply: J. B. Plotkin, J. Dushoff, H. B. Fraser
(doi:10.1038/nature03224)

Evolutionary genomics

Codon bias and selection on single genomes

Arising from: J. B. Plotkin, J. Dushoff & H. B. Fraser *Nature* **428**, 942–945 (2004)

The idea that natural selection on genes might be detected using only a single genome has been put forward by Plotkin and colleagues¹, who present a method that they claim can detect selection without the need for comparative data and which, if correct, would confer greater power of analysis with less information. Here we argue that their method depends on assumptions that confound their conclusions and that, even if these assumptions were valid, the authors' inferences about adaptive natural selection are unjustified.

The volatility analysis of Plotkin *et al.*¹ rests on the observation that synonymous codons often differ in the number of mutations that take them to different amino acids. For instance, CGA and AGA both code for arginine, but differ in the number of amino acids that are one mutation away (4 out of 8 for CGA and 6 out of 8 for AGA; AGA therefore has a higher volatility). Their expectation is that proteins that have undergone more amino-acid substitutions will have more highly volatile codons.

To test each gene for selection by looking at relative codon volatility, Plotkin *et al.* construct the null by drawing alternative synonymous codons from a distribution parameterized by genome-wide codon frequencies, with a view to identifying genes that are atypically rich or poor in volatile codons and which they believe represent rapidly and slowly evolving genes, respectively.

A premise of this approach is that codon usage does not vary in a consistent way from gene to gene. If codon bias is related to volatility in any way — for instance, if CGA is preferentially used over AGA in highly expressed genes — then the volatility index that the authors use is simply an alternative measure of within-genome variation in codon-usage bias. In fact, differential codon bias due to both natural selection and mutation bias results in a highly heterogeneous distribution of codon usage across multiple genomes^{2–5}.

To investigate the extent to which the authors' unmet assumption will affect estimation of volatility, we conducted two analyses. First, we examined the correlation between volatility *P* values and the codon adaptation index (CAI), a common measure of optimal codon bias, for every gene in the *Saccharomyces cerevisiae* genome (Fig. 1a). We use *S. cerevisiae* because the optimal codons are known⁶ and because nucleotide divergence can be estimated from a closely related out-group (*S. paradoxus*). As expected, we found a strong correlation between CAI and volatility, with CAI explaining a much larger proportion of the variance in volatility than the standard comparative measure of selection, dN/dS.

That codon usage bias — measured by either CAI or volatility — correlates with selective constraint is well known and unsurprising^{7,8}.

Second, because CAI measures only one of many known biases in codon usage, to examine the general effects of differential codon usage we randomly assigned volatility scores to each codon and then re-analysed the *Mycobacterium tuberculosis* genome according to the method of Plotkin *et al.*¹. Figure 1b shows the distribution of volatility scores for this randomized genome. This distribution has a U-shape similar to that found in the true volatility distribution for *M. tuberculosis*, and there is high similarity in the set of genes that reside in the tails of both distributions ($P < 10^{-15}$; Fig. 1b); these outliers show dissimilar codon usage from the majority of the genes, regardless of the measure used. That a random assignment of codon volatilities recovers the same outlier genes as the true values suggests that volatility itself has little to do with the observed distribution. Our analyses indicate that volatility may be another measure of codon bias.

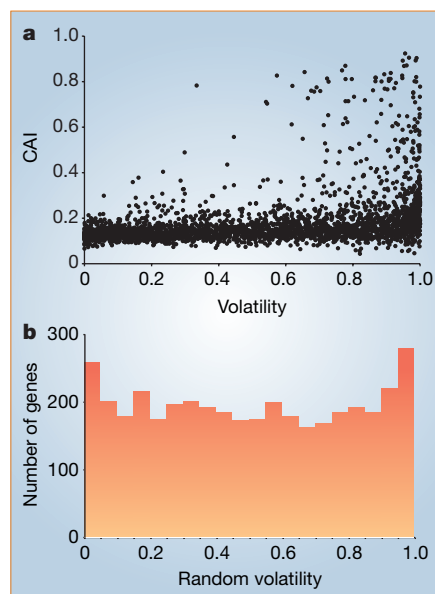


Figure 1 Codon bias explains volatility. **a**, Relationship between the codon adaptation index (CAI) and the volatility *P* values in the *Saccharomyces cerevisiae* genome. An analysis of variance with CAI and dN/dS (calculated from a comparison with orthologs in *S. paradoxus*) as main effects reveals that CAI (*F* ratio = 322, $P < 0.0001$) explains more than twice as much of the variation in volatility as dN/dS (*F* ratio = 153, $P < 0.0001$). **b**, Distribution of random volatilities in the *Mycobacterium tuberculosis* genome. We randomly assigned each codon a volatility score between 0 and 1 and calculated the volatility *P* values for all the genes in the *M. tuberculosis* genome using these new values, following the method of Plotkin *et al.*¹. The proportion of the 78 genes in the 1% tails of the volatility distribution that are shared between the random and true volatility distributions is 35% (27 genes in both). This overlap is highly non-random ($P < 10^{-15}$).

The authors do not provide formal reasoning to explain why the volatility statistic should correlate with selective constraint. If their method is grounded in standard, single-locus multiple-allele models, their result is trivial: in such models, assigning an allele lower fitness will deterministically lower its frequency in a population. And because volatility only applies to four codon families, three of which contain synonymous codons that cannot be reached by a single mutation, any such population-genetics model involving volatility must violate basic assumptions.

The authors claim both that their method does not rely on some of the strongest assumptions of comparative analyses¹ and that their unstated model assumes a population at mutation–selection balance. However, comparative methods for detecting natural selection (see ref. 9, for example) do not require populations in mutation–selection balance. Codon-based comparative methods do require a mutational process at stationarity⁹, but so do Plotkin *et al.*: if they do not assume mutational stationarity, their logic does not work.

Consider pseudogenes: volatility scores of these genes will reflect past processes, not current selection, until they reach a stationary state. Volatility is only expected to be associated with amino-acid turnover as a genome approaches codon-usage stationarity, but it is not clear how to test the assumption that a genome has reached such a state.

The distribution of negative selection across a genome is unknown and, because of this — even if all the assumptions of the model have been met — one cannot say with certainty that genes in the tails of the distribution are under increased positive selection or decreased negative selection. Simply because there are a handful of genes thought to be under positive selection present in these tails does not prove that all or even most of these genes are. Although Plotkin *et al.* provide a caveat to inferences about positive selection in their penultimate paragraph¹, this is undermined by the seven preceding claims of positive selection in the paper. Comparative analyses such as dN/dS do require data from multiple taxa, but they provide a clear statistical criterion for detecting positive selection.

Matthew W. Hahn, Jason G. Mezey, David J. Begun, John H. Gillespie, Andrew D. Kern, Charles H. Langley, Leonie C. Moyle

Center for Population Biology and Section of Evolution and Ecology, University of California, Davis, California 95616, USA

e-mail: mwhahn@ucdavis.edu

doi:10.1038/nature03221

- Plotkin, J. B., Dushoff, J. & Fraser, H. B. *Nature* **428**, 942–945 (2004).
- Akashi, H. *Genetics* **136**, 927–935 (1994).
- Chiappello, H. *et al. Gene* **209**, GC1–GC38 (1998).
- Coghlan, A. & Wolfe, K. H. *Yeast* **16**, 1131–1145 (2000).
- Marais, G., Mouchiroud, D. & Duret, L. *Proc. Natl Acad. Sci. USA* **98**, 5688–5692 (2001).
- Ikemura, T. *Mol. Biol. Evol.* **2**, 13–34 (1985).

7. Sharp, P. M. *J. Mol. Evol.* **33**, 23–33 (1991).
 8. Akashi, H. *Curr. Opin. Gen. Dev.* **11**, 660–666 (2001).
 9. Goldman, N. & Yang, Z. *Mol. Biol. Evol.* **11**, 725–736 (1994).
- Reply: J. B. Plotkin, J. Dushoff and H. B. Fraser reply to this communication (doi:10.1038/nature03224).

Evolutionary genomics

Detecting selection needs comparative data

Positive selection at the molecular level is usually indicated by an increase in the ratio of non-synonymous to synonymous substitutions (dN/dS) in comparative data. However, Plotkin *et al.*¹ describe a new method for detecting positive selection based on a single nucleotide sequence. We show here that this method is particularly sensitive to assumptions regarding the underlying mutational processes and does not provide a reliable way to identify positive selection.

Plotkin *et al.*¹ use a measure for detecting selection known as the volatility index, whereby a codon with high volatility is more likely to have arisen by a non-synonymous mutation than a codon with low volatility; so, for high dN/dS, there should be more codons of high volatility. Positive selection should be detectable simply by examining the volatility index in a single sequence.

However, this argument is flawed because high rates of non-synonymous mutation will increase the rate of substitution both into and out of codons with high volatility. In models in which the substitution process is reversible over time, these two factors will cancel each other out, and variations in the strength of selection at the amino-acid level do not affect the expected volatility. Although most models used in studies of molecular evolution are time-reversible², the true substitution process probably is not, because of the specifics of the mutational and population-level processes.

To examine the effect of the substitution model on the volatility index, we simulated random-substitution models in which the rate of substitution between different nucleotides was sampled from a uniform random variable between zero and one. For these models, we then calculated the equilibrium frequencies of the 61 sense codons in a Markov chain model that resulted from simulations having varying synonymous and non-synonymous substitution rates. Based on the equilibrium frequencies, we could then calculate the expected value of the volatility index.

Our results indicate that the volatility index can be either an increasing or a decreasing function of dN/dS, or have a minimum or maximum at an intermediate value of this ratio (Fig. 1). We also find that the dN/dS ratio only marginally affects the volatility index — particularly for values of

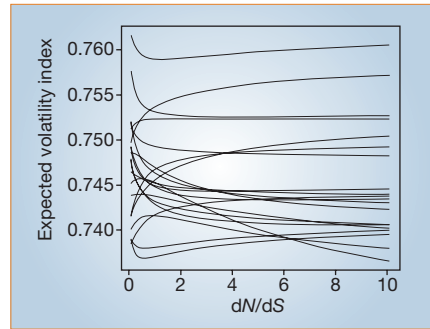


Figure 1 Expected value of the volatility index, as defined by Plotkin *et al.*¹, as a function of the dN/dS ratio for 20 random-substitution models.

dN/dS > 1. Although models can be constructed in which strong stabilizing selection on particular amino acids has a marked effect on the volatility index, there is no evidence that the volatility index captures much information regarding positive selection. Realistic models of positive selection will predict an increased rate of substitution both in and out of codons with high volatility.

What then explains the results of Plotkin *et al.*¹, in which the volatility index correlates with the rate of amino-acid substitution in comparative data and with the amount of expression? Non-random codon usage is common in most organisms, particularly in bacteria and yeast^{3–7}, and may be caused by selection for optimal codon usage and affected by variation in the nucleotide composition and other factors. In bacteria, the strength of codon-usage bias is correlated with the amount of expression^{3–5} and with the extent of amino-acid substitution^{6,7}; this may be because highly expressed genes tend to be more conserved at the amino-acid level and have more codon-usage bias than genes with low expression. The degree of amino-acid substitution might also correlate with local nucleotide frequencies because regions that differ in this respect could have different rates and patterns of mutation.

To investigate the extent to which the volatility index is sensitive to local nucleotide content, we took advantage of the fact that only codons with sixfold degeneracy or with stop codons as neighbours can contribute to the volatility index. Using all other codons we obtained independent estimates of the nucleotide frequencies. We also calculated a *P* value for a one-tailed test of increase in the frequency of a particular nucleotide by using the methodology of Plotkin *et al.*¹, but calculated only for codons that do not contribute to the volatility index.

Applying this approach to the *Plasmodium falciparum* data analysed by Plotkin *et al.*¹, the correlation coefficient between the log *P* value of the volatility index and the log *P* value associated with the percentage of thymine is 0.29. Variation in third-position nucleotide content is one of the factors explaining the distribution of volatility-

index-related *P* values in *P. falciparum*. Correlation of the volatility index with the amount of amino-acid substitution could be caused by the presence of covariates such as nucleotide frequencies, selection for optimal codon usage bias and/or expression levels.

The results of Plotkin *et al.*¹ might also be explained by variation in the amino-acid frequencies among genes. If the true evolutionary model is not time-reversible, these frequencies should influence codon usage and the volatility *P* value. Indeed, many of the amino-acid frequencies show correlation with the volatility *P* values calculated by Plotkin *et al.*¹. For example, the correlation coefficient between the frequency of glutamine and the log volatility *P* values is -0.32 . All codons for glutamine have the same volatility, but this amino acid is one mutational step away from arginine and leucine, which both affect the volatility index. The volatility index in models that are not time-reversible can therefore be affected by stabilizing selection on particular amino acids, because such selection affects the amino-acid frequency. But whether the volatility index correlates positively or negatively with such selection depends on which amino acid is the target of selection. Positive selection that increases the rate of amino-acid substitution does not have the same impact on the volatility index.

We argue that the volatility index cannot be applied to detect positive selection as it is under greater influence from other factors, such as amino-acid and nucleotide frequencies. However, the results of Plotkin *et al.*¹ should spur efforts to identify the causes of non-random codon usage in bacteria and other organisms.

Rasmus Nielsen*†, Melissa J. Hubisz†

*Centre for Bioinformatics, University of Copenhagen, Universitetsparken 15, 2100 Copenhagen, Denmark
e-mail: rasmus@binf.ku.dk

†Department of Biological Statistics and Computational Biology, Cornell University, Ithaca, New York 14853, USA

doi:10.1038/nature03222

1. Plotkin, J. B., Dushoff, J. & Fraser, H. B. *Nature* **428**, 942–945 (2004).
2. Lio, P. & Goldman, N. *Genome Res.* **8**, 1233–1244 (1998).
3. Ikemura, T. *J. Mol. Biol.* **151**, 389–409 (1981).
4. Grantham, R., Gautier, C., Gouy, M., Jacobzone, M. & Mercier, R. *Nucl. Acids Res.* **9**, 43–74 (1981).
5. Grosjean, H. & Fiers, W. *Gene* **18**, 199–209 (1982).
6. Sharp, P. M. *J. Mol. Evol.* **33**, 23–33 (1991).
7. Akashi, H. & Gojobori, T. *Proc. Natl Acad. Sci. USA* **99**, 3695–3670 (2002).

Reply: J. B. Plotkin, J. Dushoff and H. B. Fraser reply to this communication (doi:10.1038/nature03224).

Evolutionary genomics

Codon volatility does not detect selection

Plotkin *et al.*¹ introduce a method to detect selection that is based on an index called codon volatility and that

uses only the sequence of a single genome, claiming that this method is applicable to a large range of sequenced organisms¹. Volatility for a given codon is the ratio of non-synonymous codons to all sense codons accessible by one point mutation. The significance of each gene's volatility is assessed by comparison with a simulated distribution of 10^6 synonymous versions of each gene, with synonymous codons drawn randomly from average genome frequencies. Here we re-examine their method and data and find that codon volatility does not detect selection, and that, even if it did, the genomes of *Mycobacterium tuberculosis* and *Plasmodium falciparum*, as well as those of most sequenced organisms, do not meet the assumptions necessary for application of their method.

Because of codon table regularity, 16 of 20 codon families have the same volatility for each synonymous codon. The authors' assertion that there are 22 codons that have at least one synonym with a different volatility¹ obscures the fact that these 22 codons represent only four amino acids: the fourfold-degenerate glycine codons and the sixfold-degenerate arginine, leucine and serine codons. Therefore, even if the method were

valid, volatility neglects any selection acting on the remaining codons. Figure 1 shows that the volatility P values of Plotkin *et al.*¹, which are purported to detect selection, are largely a measure of the codon usage of serine, with smaller contributions from glycine, arginine and leucine. A codon-usage index dominated by a single family is not generally useful for studying codon bias, much less selection — unless the genic signature of natural selection is merely serine-codon usage.

The theoretical basis of the method of Plotkin *et al.*¹ is unclear: either it depends critically on the only theoretical justification offered by the authors, a model by van Nimwegen *et al.*², or it is unfounded. Assuming the null model is that of Nimwegen *et al.*, it requires that all synonymous codons are of equal fitness, that all members of a family are mutually accessible by a path of neutral mutation, and that the product of the effective population size and mutation rate, $N_e\mu$, is much greater than one. But these requirements are respectively violated in that selection for codon usage in highly expressed genes is common³; the twofold-degenerate serine codons cannot access the fourfold-degenerate codons without passing through a non-synonymous intermediate; and most sequenced genomes

have an estimated $N_e\mu \ll 1$ (ref. 4). Specifically, $N_e\mu$ for pathogens is constrained by repeated bottlenecks caused by transmission. In fact, the highest estimated $N_e\mu$ values for both *P. falciparum* and *M. tuberculosis* are of the order of 10^{-3} or less^{5,6}.

In Table 1 of Plotkin *et al.*¹, the authors present ten genes (including five of putative function, not identified as such, and one hypothetical protein) that they say show the most significant volatility P values and that therefore “show the strongest signs of positive selection”, although they ignore 27 other genes (see their supplementary information) with P values indistinguishable from those in their table. These include a putative cell-cycle control protein, histone deacetylase, and DNA-directed RNA polymerase — highly conserved genes not expected to show signs of recent positive selection. Furthermore, genes encoding transfer RNA and ribosomal RNA are included in their analysis and assigned volatility P values, although these genes do not have codons.

The reality at present is that the community must continue to rely on other methods to detect natural selection.

Ying Chen*, J. J. Emerson*, Todd M. Martin†

*Department of Ecology and Evolution, and
†Committee on Genetics, University of Chicago,
Chicago, Illinois 60637, USA

doi:10.1038/nature03223

1. Plotkin, J. B., Dushoff, J. & Fraser, H. B. *Nature* **428**, 942–945 (2004).
2. van Nimwegen, E., Crutchfield, J. P. & Huynen, M. *Proc. Natl Acad. Sci. USA* **96**, 9716–9720 (1999).
3. Akashi, H. *Curr. Opin. Genet. Dev.* **11**, 660–666 (2001).
4. Lynch, M. & Conery, J. S. *Science* **302**, 1401–1404 (2003).
5. Hughes, A. L. & Verra, F. *Proc. R. Soc. Lond.* **268**, 1855–1860 (2001).
6. Sreevatsan, S. *et al. Proc. Natl Acad. Sci. USA* **94**, 9869–9874 (1997).

Reply: J. B. Plotkin, J. Dushoff and H. B. Fraser reply to this communication (doi:10.1038/nature03224).

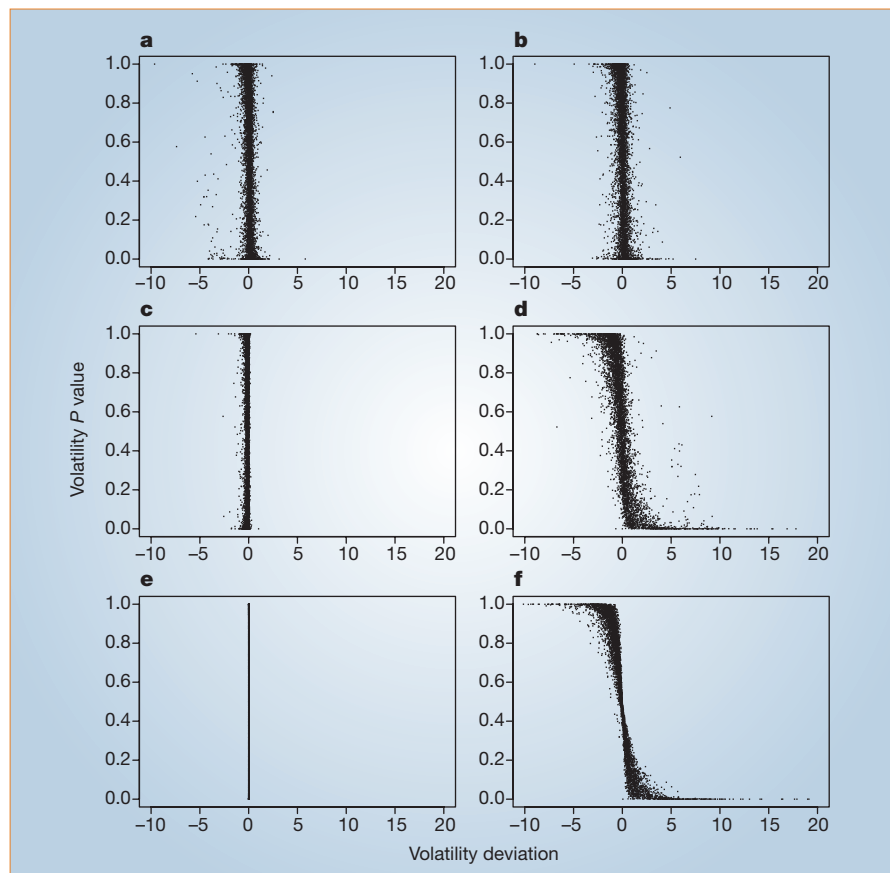


Figure 1 Volatility P value plotted against volatility deviation. For each gene in the *Plasmodium falciparum* genome, the relationship between the volatility P values of Plotkin *et al.*¹, which reportedly consider all codon contributions, and the difference between the observed volatility contributions for given codons and the volatility contributions expected by their genome average frequency is shown. Serine is the only amino acid for which the tails of the statistic are significantly associated with extremes in the P value. This analysis was repeated for each gene in the *Mycobacterium tuberculosis* genome (data not shown). **a**, Arginine; **b**, leucine; **c**, glycine; **d**, serine; **e**, all other families, excluding arginine, leucine, glycine and serine; and **f**, all 20 families.

Plotkin et al. reply — The criticisms of our results¹ by Hahn *et al.*², Nielsen and Hubisz³, and Chen, Emerson and Martin⁴ fall into three categories: the formal justification for our method, the potential for confounding factors, and the interpretation of our empirical results.

In response to assertions^{2,4} that we do not formally discuss why codon volatility should reflect selection pressures on proteins, our method is grounded in the standard multi-allele model of population genetics^{5,6}. Hahn *et al.* note that assigning an allele lower fitness will deterministically lower its frequency in a population², but this fundamentally misunderstands why volatility reflects selection: even if all synonymous codons for an amino acid are assigned equal fitness, selection at the amino-acid level will bias the relative frequencies of synonymous codons.

Consider, for example, a site under negative selection for arginine. Assigning equal fitness to arginine codons and lower fitness to all other codons, the multi-allele model⁵ indicates that a less volatile arginine

codon will occur with higher equilibrium frequency⁶. The expected volatility at such a site depends monotonically on the strength of selection for arginine⁶. In a finite population at steady state, the standard Fisher–Wright model⁷ yields the same result, although the strength of the volatility signal depends upon the population size⁶. We agree with Hahn *et al.* (and never claimed otherwise) that this logic implicitly assumes stationarity, as is likewise assumed by all comparative methods for estimating selection^{2,8}.

Nielsen and Hubisz³ question whether volatility can detect selection because, they claim, “variations in the strength of selection at the amino-acid level do not affect the expected volatility”. This claim is true only under simplified models of sequence evolution that ignore population variability⁸. Such models (used to produce their Fig. 1) are reasonable approximations when comparing divergent lineages⁸, but they fail to detect the effect of amino-acid selection on synonymous codon usage.

There is an intuitive explanation for this. Such models consider a single sequence that is assumed to represent the dominant genotype in a population at any time. Mutation and selection are modelled simultaneously by adjusting the transition rates between codon states⁸. Under the standard assumption of time reversibility^{3,8}, the equilibrium rate of transitions into a codon state must equal the rate out of that state: therefore, synonymous codons occur with the same probability regardless of the strength of negative selection³.

By contrast, in a more realistic Fisher–Wright model of a replicating population, the equilibrium rate of mutations away from a codon is not compensated by an equal rate of mutations back into that codon: many lineages that acquire a non-synonymous mutation are removed from the population by selection, before back-mutation. As a result, codons with a greater chance per unit time of non-synonymous mutation (that is, more volatile codons) occur with lower equilibrium frequency⁶. Thus, models that ignore population variability^{3,9} do not adequately describe the patterns of synonymous codon usage.

Chen *et al.*⁴ note that our treatment of serine violates van Nimwegen’s model¹⁰. But our method is grounded more generally in the standard multi-allele model⁵, whose assumptions are not violated by the arrangement of serine codons. In fact, serine is particularly informative for positive selection⁶, which may explain why more signal is derived from serine⁴.

These authors argue⁴ that $N_e\mu$ is too small to support a volatility signal of negative selection for *Plasmodium falciparum* or *Mycobacterium tuberculosis*. But the effective population sizes of microorganisms are debatable. Estimates of $N_e\mu$ can vary by four orders of magnitude and depend strongly on assumptions

about population structure¹¹ and homogeneity of mutation rates¹². More important, the population size estimated from neutral polymorphism data⁴ is generally not the same as the population size relevant to volatility: common phenomena such as transient hypermutators in microbes can vastly improve the power of volatility to detect negative selection, even when the typical neutral site heterozygosity is much less than one⁶. Moreover, even if the relevant effective population size is small, there is a signal (of the order of $N_e\mu$ per site) that distinguishes negative selection from neutrality⁶. Last, concerns about population size do not apply in the context of positive selection: a more volatile codon is much more likely to occur at a site that has undergone a positively selected sweep, regardless of the population size⁶.

The authors^{2–4} are concerned that the volatility signal may be compromised by other sources of differential selection on synonymous codons across a genome — a point that we also highlighted¹ and which likewise applies to dN/dS. The observation² that volatility P values correlate with the codon adaptation index (CAI) in yeast does not invalidate our method of estimating selection, especially given that dN/dS also correlates with CAI, to the same extent⁶. Moreover, volatility P values are significantly correlated with dN/dS even after controlling for CAI (partial correlation $P < 10^{-33}$) — indicating that the two measures yield fairly consistent estimates of selection, even aside from any signal inherent in CAI. The observation³ that volatility correlates with amino-acid or nucleotide content does not discredit its ability to detect selection; analogous correlations are found for dN/dS.

Hahn *et al.*² note that several *M. tuberculosis* genes exhibit extraordinary codon usage and will typically appear in one or the other tail of the distribution when assigning “random volatility”². Although these outliers may influence the overall shape of the distribution, it is critical to recognize that “random volatility”² does not assign the same genes to the same tails as volatility¹ does. When applied to *M. tuberculosis*, random volatility is not correlated with the rate of protein evolution ($P = 0.23$), nor does it show increased values among the antigenic PE/PPE proteins ($P = 0.8$), nor depressed values among those genes required for bacterial growth ($P = 0.4$) or among those genes conserved between species ($P = 0.5$). By contrast, our volatility measure shows highly significant correlations ($P < 10^{-6}$) in all four of these cases¹ — indicating that volatility is not simply an alternative metric of codon bias, but rather a measure that co-varies with selection pressures on *M. tuberculosis* proteins, as expected⁶.

Chen *et al.*⁴ note that volatility detects a signal from four amino acids, comprising about 25% of sites in a typical gene, and that more signal arises from serine in *P. falciparum*.

They suggest that a signal obtained from so few sites would be insufficient to detect selection, but provide no evidence to support this claim. By contrast, we present strong empirical evidence that volatility P values reflect known selection pressures on proteins¹. Moreover, comparative analyses of selection often rely on signals obtained from fewer than 5% of sites^{13,14}.

Chen *et al.*⁴ contend that volatility fails to detect selection because they find three highly volatile *P. falciparum* genes that they do not expect to experience positive selection, based on their putative biological functions. Even assuming that the putative functional annotations are correct, Chen *et al.* present no evidence that these three genes experience strong negative selection. It is possible (owing to gene duplication, for example) that members of a highly conserved gene family experience relaxed negative or even positive selection¹⁵; in fact, these three genes each have at least one paralogue in *P. falciparum*. In any case, even if these three genes do actually experience negative selection, a handful of false positives among a screen of over 5,000 genes would hardly invalidate our method of estimating relative selection pressures across a genome.

Hahn *et al.* point out that it cannot be said with certainty that genes in the tails of the distribution are under increased positive or decreased negative selection². We agree, and continue to emphasize that volatility P values are intrinsically relative: we still cannot conclude that any gene is under positive selection in an absolute sense and can only conclude that some genes are under more positive, or less negative, selection than others¹.

J. B. Plotkin*, J. Dushoff†‡, H. B. Fraser§

*Harvard Society of Fellows, 7 Divinity Avenue, Cambridge, Massachusetts 02138, USA
e-mail: jplotkin@fas.harvard.edu

†Department of Ecology and Evolutionary Biology, Princeton University, Princeton, New Jersey 08540, USA

‡Fogarty International Center, National Institutes of Health, Bethesda, Maryland 20892, USA

§Department of Molecular and Cell Biology, University of California, Berkeley, California 94720, USA

doi:10.1038/nature03224

- Plotkin, J. B., Dushoff, J. & Fraser, H. B. *Nature* **428**, 942–945 (2004).
- Hahn, M. W. *et al.* *Nature* doi:10.1038/nature03221 (2004).
- Nielsen, R. & Hubisz, M. J. *Nature* doi:10.1038/nature03222 (2004).
- Chen, Y., Emerson, J. J. & Martin, T. M. *Nature* doi:10.1038/nature03223 (2004).
- Haldane, J. B. S. *Proc. Camb. Phil. Soc.* **23**, 838–844 (1927).
- Plotkin, J. B. *et al.* Preprint at <http://arxiv.org/abs/q-bio/0410013> (2004).
- Wright, S. *Genetics* **16**, 97–159 (1931).
- Goldman, N. & Yang, Z. *Mol. Biol. Evol.* **11**, 725–736 (1994).
- Dagan, T. & Graur, D. *Mol. Biol. Evol.* doi:10.1093/molbev/msi033 (2004).
- van Nimwegen, E., Crutchfield, J. P. & Huynen, M. *Proc. Natl Acad. Sci. USA* **96**, 9716–9720 (1999).
- Berg, O. *Genetics* **142**, 1379–1382 (1996).
- Tajima, F. *Genetics* **143**, 1457–1465 (1996).
- Fleischmann, R. D. *et al.* *J. Bacteriol.* **184**, 5479–5490 (2002).
- Clark, A. G. *et al.* *Science* **302**, 1960–1963 (2004).
- Katz, L. A. *et al.* *Mol. Biol. Evol.* **21**, 555–562 (2004).

nature

A Nature Publishing Group Supplement



year of physics a celebration

In 1905, Albert Einstein submitted five papers for publication in *Annalen der Physik*, covering three topics: the photoelectric effect, brownian motion, and the special theory of relativity. Although diverse in subject matter, these contributions are landmarks in their field — and testament to Einstein's genius. To honour their centenary, 2005 has been designated 'World Year of Physics'. *Nature* joins the celebrations with the publication of this special supplement.

We offer you a collection of essays and review articles inspired by the themes of Einstein's 1905 work: thermodynamics, quantum physics and, of course, relativity.

Looking forward, these themes are represented here in discussions of the fluctuation–dissipation theorem, quantum criticality and quantum optics, symmetry-breaking and the Higgs boson, and the frontiers of cosmology. But we also look back to the events of 1905 and the wider context of Einstein's *annus mirabilis*, and consider the distinctiveness and durability of his image. To round off the celebration, physicists including Steven Weinberg and Roger Penrose give their take on the 'theory of everything' that famously eluded Einstein.

And there's one more contribution to World Year of Physics that we would like to mention: in October 2005, Nature Publishing Group will launch *Nature Physics*, a new monthly journal. Alongside *Nature*, *Nature Physics* will publish research, commentary and analysis across the entire spectrum of physics, pure and applied.

We hope you enjoy 'year of physics'!

Alison Wright Editor, *Nature Physics*
Karl Ziemelis Physical Sciences Editor, *Nature*
Leslie Sage Senior Editor, *Nature*
Karen Southwell Senior Editor, *Nature*

- commentary**
215 1905 and all that
J. Stachel
- essay**
218 Einstein as icon
J. D. Barrow
- concept**
221 Brownian motion
G. Parisi
- progress articles**
222 In and out of equilibrium
J. Kurchan
- 226 Quantum criticality**
*P. Coleman &
A. J. Schofield*
- review articles**
230 Happy centenary, photon
*A. Zeilinger, G. Weihs,
T. Jennewein &
M. Aspelmeyer*
- 239 In search of symmetry lost**
F. Wilczek
- 248 The state of the Universe**
P. Coles
- endgame**
257 A theory of everything?

Cover illustration

Albert Einstein from the Series 'Ten Portraits of Jews of the Twentieth Century' by Andy Warhol, 1980 (Andy Warhol Foundation/CORBIS)

Editor, *Nature*: **Philip Campbell**
Production Editor: **Lucy Odling-Smee**
Art Editor: **Martin Harrison**

Supplement Publisher: **Sarah Greaves**
Production: **Sue Gray**
Marketing: **Claire Aspinall**
Sponsorship: **Claire Hines**

Layouts: **Nadia Rooney**
Diagrams: **Nik Spencer**
Michael Ayres

1905 and all that

John Stachel

How Einstein claimed his place in the changing landscape of physics during his *annus mirabilis*.

At the end of 1904, a reader who happened to glance over the contributions to the *Annalen der Physik* (the premier German physics journal) of one Albert Einstein, an obscure clerk at the Swiss patent office, would have found just five articles — the first published in 1901, when its author was 22 years old. These articles constituted Einstein's entire published *oeuvre*, and none was sufficiently distinguished for our hypothetical reader to anticipate the nature or significance of Einstein's next five papers^{1–5}, all submitted to the *Annalen* in 1905, his *annus mirabilis*.

Someone privy to Einstein's correspondence or in close contact with him — his fellow physics student and wife Mileva Marić or his old friend and fellow patent office clerk Michele Besso, for example — would have been better prepared. These people knew that, at least since his student days at the Swiss Federal Polytechnic (1896–1900), the young Einstein had been concerned with the foundations of theoretical physics. He had been probing the edifice erected by his predecessors in this field for its strengths and weaknesses, and was already beginning to suggest modifications of some core assumptions⁶.

Newton's legacy in question

During the nineteenth century, the mechanistic world-view — based on Isaac Newton's formulation in the *Principia* (1687) of the kinematics and dynamics of corpuscles of matter, and crowned by his stunningly successful theory of gravitation — was challenged first by the optics, then by the electrodynamics of moving bodies. By the mid 1800s Newton's corpuscular theory of light was no longer tenable. To explain Snell's law of refraction, this theory assumed that light corpuscles speed up on encountering a medium of higher refractive index. But in 1849, Léon Foucault and Hippolyte Fizeau showed that, in fact, light slowed down, as predicted by the rival wave theory espoused by Newton's contemporary Christiaan Huygens. The problem now was to fit the wave theory of light into the newtonian picture of the world. Indeed, the ether — the medium through which light waves were assumed to propagate in the absence of ordinary, ponderable matter — seemed to provide a physical embodiment of Newton's absolute space. But elucidating the relation between ether and ponderable matter presented grave problems: did moving matter drag the ether with it — either totally or partially — or did the ether remain immobile? It proved impossible to reconcile the consequences of any of these hypotheses with all the experimental results on the optics of moving bodies. By the last third of the nineteenth century, many physicists were acutely aware of this problem⁷.

By 1865, James Clerk Maxwell had showed that light could be interpreted as wave-like oscillations of the electric and magnetic fields, obeying what we now call the Maxwell equations for these fields. It was realized that the optical problems were only a special case of similar problems in reconciling the electrodynamics of moving bodies with newtonian kinematics and dynamics. Towards the end of the century, however, Hendrik Antoon Lorentz seemed to

overcome all these problems through his interpretation of Maxwell's equations. Lorentz assumed that the electromagnetic ether is entirely immobile, in which case there would be no dragging of the ether.

Although in newtonian mechanics it is impossible to distinguish any preferred inertial frame (this result is often referred to as the galileian principle of relativity), at first the situation seemed different for electrodynamics and optics. The rest frame of the ether provided a preferred inertial frame, and motion through it should have been detectable. Yet all attempts to detect the translational motion of the Earth through the ether by means of optical, electrical or magnetic effects consistently failed. Lorentz succeeded in explaining why: according to his theory, no such effect should be detectable by any experiment sensitive to first order in (v/c) , where v is the speed of the moving object through the ether and c is the speed of light in that medium. Until the 1880s, no experiment with greater sensitivity had been performed, and Lorentz's explanation of the failure of all previous experiments was a crowning achievement of his theory.

Newton's mechanics now seemed to have successfully met the challenge of optics and electrodynamics. But the seeds of its downfall had already been planted. Lorentz's explanation led him to introduce a transformation from newtonian absolute time to a new time variable in each inertial frame moving through the ether. As the relation between absolute time and this time varied from place to place in each inertial frame, Lorentz called this new variable the 'local time' of that frame, regarding the local time as a purely formal expression. But Henri Poincaré, the great mathematician who concerned himself extensively with problems of physics, was able to give a physical interpretation of this time variable within the context of newtonian kinematics: it is the time that clocks at rest in a frame moving through the ether would read if they were synchronized using light signals, without taking into account the motion of that frame. This was an important hint that the problems of the electrodynamics and optics of moving bodies were connected with the concept of time. But, as we shall see, it was Einstein who made the final break with the concept of absolute time, by asserting that the local time of any inertial frame is as physically meaningful as that of any other because there is no absolute time with which they can be compared.

But by the time Lorentz's theory was completed, it was in trouble. Albert Michelson, a physicist specializing in optical interferometry, performed an experiment to detect motion through the ether that was sensitive to order $(v/c)^2$ and, according to Lorentz's theory, this experiment should have succeeded. But it did not. So Lorentz was forced to start patching up his theory by introducing a transformation to new spatial coordinates in a moving inertial frame. (He was ultimately also forced to modify his definition of the local time, and it is this modified transformation law that agrees formally with Einstein's time-transformation law.) For Lorentz's spatial transformation to eliminate the predicted second-order effect, Lorentz interpreted it as corresponding

5. Über die von der molekularkinetischen Theorie der Wärme geforderte Bewegung von in ruhenden Flüssigkeiten suspendierten Teilchen;
von A. Einstein.

In dieser Arbeit soll gezeigt werden, daß nach der molekularkinetischen Theorie der Wärme in Flüssigkeiten suspendierte Körper von mikroskopisch sichtbarer Größe infolge der Molekularbewegung der Wärme Bewegungen von solcher Größe ausführen müssen, daß diese Bewegungen leicht mit dem Mikroskop nachgewiesen werden können. Es ist mir gelungen, die hier zu behandelnden Bewegungen mit der „Brownischen Molekularbewegung“ identisch zu machen. Die erreichbaren Angaben über letztere sind so genau, daß ich mir hierüber kein Urteil bilden kann.

Wenn sich die hier zu behandelnden Bewegungen für sie zu erwartenden Gesetzmäßigkeiten unterwerfen läßt, so ist die klassische Theorie der mikroskopisch unterscheidbaren Teilchenbewegungen anzusehen und es ist zu erwarten, daß die Atomgröße möglichst genau dieser Bewegung entspricht. Die wesentliche Wiegende der Wärme ist die Kontraktion der Körper. We now call this effect the Lorentz contraction, even though, thanks to Einstein, it is regarded as the purely kinematical effect of comparing lengths in two inertial frames, rather than a dynamical contraction due to motion through the ether.

The kinetic–molecular challenge

In spite of the problems posed by electrodynamics, the mechanical picture of the world was far from exhausting its potential at the end of the nineteenth century. A series of brilliant theoretical studies by Maxwell, Rudolf Clausius and Ludwig Boltzmann began to show how the empirically established laws of thermodynamics could be explained in terms of kinetic–molecular models of matter, first for gases and later for liquids and solids. Many properties of materials, such as thermal conductivity and viscosity, could also be explained by such a model. But many aspects of this kinetic–molecular theory of heat remained obscure. For example, how could the time-reversible laws of mechanics governing the behaviour of material particles give rise to the time-irreversible behaviour of bulk matter, summarized in the second law of thermodynamics?

The ‘energeticists’ — a group of physicists and especially physical chemists — challenged the entire kinetic–molecular theory of heat, demanding that all physics be based on the macroscopic concept of energy and casting doubt on the kinetic–molecular hypothesis itself. They held that the laws of thermodynamics were strictly correct. In kinetic–molecular theory, however, the laws give the extremely probable results of some averaging procedure over an immensely large number of molecules in bulk matter. Demonstrating microscopic violations of the laws of thermodynamics was thus a crucial challenge for the proponents of kinetic–molecular theory.

The newly-developed laws governing the behaviour of electromagnetic radiation in thermal equilibrium also presented a challenge. (This electromagnetic radiation is called ‘cavity radiation’ or *Hohlraumstrahlung* in German because it is the state that such radiation would reach were it confined to a cavity with walls at a fixed temperature, and it is called ‘black-body radiation’ in English because it is the

type of radiation that would be emitted by a perfectly absorbing, and hence black-appearing, body.) The total energy of the black-body radiation contained in a unit volume and its dependence on temperature could be explained by the laws of thermodynamics and electrodynamics, but the distribution of this radiant energy over different frequencies defied classical calculation. If such a calculation were attempted consistently, it gave the absurd result that the total energy in a unit volume must be infinite, no matter how low the temperature of the radiation. By the turn of the century, experiment had produced a well-behaved frequency distribution formula for the energy radiated by a black body, in the form of Planck’s law. The challenge then was to explain Planck’s law theoretically. Planck provided such a derivation in 1900, but its physical significance for the structure of matter and radiation remained obscure.

Enter Einstein

Einstein set about tackling all the problem areas mentioned above. Far from being an unqualified revolutionary, his first papers, during 1904, were concerned with the further development of classical mechanics. In particular, Einstein focused on thermodynamics and the kinetic–molecular theory of heat, and attempted to fill some gaps that he saw in the theory (although he later admitted that some of these gaps were in his knowledge of the existing literature). By 1905, Einstein was intent on demonstrating both the reality of molecules, by showing how their size could be estimated from the effect of a suspension of small particles in a fluid on the viscosity of that fluid, and the necessity of the kinetic–molecular theory, by showing that microscopic violations of the laws of thermodynamics do exist. He not only proved that fluctuation phenomena — inconsistent with purely macroscopic thermodynamics but easily explained by the kinetic–molecular theory of heat — exist, but also realized that they had been observed decades earlier as the brownian motion of pollen grains in water².

Einstein derived formulae for the change in viscosity of a solvent as a function of the fraction of solute present; for the diffusion coefficient of small particles suspended in a liquid; and for the variation of the mean free path of a brownian particle as a function of the time interval during which it is observed. For the change in solvent viscosity, the original calculation was wrong, proving that no one is perfect; only after discrepancies with experiment came to light did Einstein find the error and publish a corrected result in 1911. The last result, modelling brownian motion, is the first example of the successful theoretical treatment of a stochastic process. Because of their practical importance — in fields as varied as the study of aerosol particles, the properties of milk, and semiconductor physics — Einstein’s studies on molecular size⁵ and brownian motion² are the most cited of his 1905 *oeuvre*.

By this time, some physicists (such as Max Abraham) advocated abandoning the mechanical world-view in favour of an electromagnetic world-picture based on Maxwell’s theory; others (including Planck) wanted to tinker with classical mechanics and Maxwell’s theory — in particular, with the laws governing the exchange of energy between mechanical oscillators and the electromagnetic field. But Einstein’s profound concern with the nature of thermal radiation led him to the conclusion that neither classical mechanics nor Maxwell’s electrodynamics could survive intact. Both would have to be modified to take account of Planck’s discovery in 1900 of the quantum of action, h . To explain the laws of thermal radiation and the exchange of energy between matter and radiation, quantum theories of matter and radiation would be required. It was this aspect of his work that Einstein characterized as “very revolutionary”⁸.

In his first published paper¹ of 1905, Einstein suggested that electromagnetic radiation at high frequencies could be thought of as

composed of 'light quanta', and thereby founded the field of quantum optics; but this was a step considered so radical at the time that few followed him in taking it. Planck himself was not convinced and almost a decade later, when recommending Einstein for a position in the Prussian Academy of Sciences, he still felt it necessary to excuse Einstein for this apparent mistake. (It was only the discovery of the Compton effect in 1923 that made the photon concept respectable.)

By 1907, Einstein had applied the quantum hypothesis to crystalline solids. By treating a crystal as an ensemble of particles oscillating about their equilibrium position with quantized energies, he was able to explain the anomalously low specific heat of such solids at low temperatures. In fact, it was the successful experimental verification of Einstein's formula for specific heat that first brought the quantum theory to the attention of most physicists⁹.

Classical physics completed

Einstein never regarded his work on resolving the apparent conflict between classical mechanics and electrodynamics^{3,4} — which led to what we now call the special theory of relativity — as revolutionary in the same sense as his work on the quantum hypothesis. He saw it rather as the culmination and completion of classical physics. His decade-long confrontation with the problems of the electrodynamics of moving bodies finally led to a breakthrough in 1905. Einstein realized that, at the classical (non-quantum) level, all the contradictions between mechanics and electrodynamics could be removed by recognizing the need for — and the feasibility of — a radical modification of the newtonian concept of absolute time. This modification led to a new kinematical foundation for all of physics.

Einstein based this modification on a new interpretation of the Lorentz–Poincaré local time. Instead of regarding this as an 'apparent but false' time for each inertial frame, as compared to the one true, absolute and universal time of Newton, Einstein saw it as a possible definition of time that could be introduced in each inertial frame, in such a way that the vacuum speed of light is the same in every inertial frame (he dropped all reference to the privileged ether frame). Note the subtle but crucial difference. Poincaré had interpreted the local time as that given by clocks at rest in a frame moving through the ether when synchronized as if — contrary to the basic assumptions of newtonian kinematics — the speed of light were the same in all inertial frames. Einstein dropped the ether and the 'as if': one simply synchronized clocks by the Poincaré convention in each inertial frame and accepted that the speed of light really is the same in all inertial frames when measured with clocks so synchronized.

Of course, the galileian law of addition of velocities can no longer hold, so a new kinematics was required. Einstein showed how to set up such a kinematics, which we now call special-relativistic. The principle of relativity (that is, the democracy of all inertial frames), which in newtonian theory held only for mechanical phenomena, now holds for all phenomena, in particular for all optical and electromagnetic phenomena. So negative results, such as that of Michelson, do not call for a dynamical explanation (as they did in Lorentz's version of Maxwell's theory). They are simply a consequence of the universal validity of the principle of relativity as an empirical generalization; just as the impossibility of all perpetual motion machines follows from the laws of thermo-dynamics.

The theoretical reason for the earlier problems in reconciling mechanics and electrodynamics became clear. Maxwell's laws of electrodynamics are invariant under the space-time transformations of the new kinematics, now called the Lorentz transformations; the laws of mechanics were invariant under the space-time transformations of the old kinematics, usually called Galilei transformations. So classical mechanics had to be

replaced by a new special-relativistic mechanics for particles and continuous media (rigid bodies proved incompatible with the special theory), and this was developed by Einstein, Planck, Max Born, Max von Laue and others over the next decade.

The new concepts of space and especially of time that were inherent in the new kinematics proved difficult for many physicists to assimilate — and indeed still are for many. Yet Einstein's most revolutionary revision of these concepts still lay ahead, when he turned in 1907 to the problem of fitting gravitation into his new kinematic framework. He soon convinced himself — although, again, it was a long time before others were persuaded — that this could not be done within the bounds of the special theory of relativity, and he embarked on another decade-long odyssey that led him to the general theory. But that is a story for another occasion. ■

John Stachel is at the Department of Physics and Director of the Center for Einstein Studies, Boston University, Boston, Massachusetts 02215, USA

e-mail: stachel@physics.bu.edu

He is the editor of Einstein's *Miraculous Year: Five Papers That Changed the Face of Physics*. A new edition will appear in May 2005.

1. Einstein, A. *Ann. Phys. (Leipz.)* 17, 132–148 (1905).
2. Einstein, A. *Ann. Phys. (Leipz.)* 17, 549–560 (1905).
3. Einstein, A. *Ann. Phys. (Leipz.)* 17, 891–921 (1905).
4. Einstein, A. *Ann. Phys. (Leipz.)* 18, 639–641 (1905).
5. Einstein, A. *Ann. Phys. (Leipz.)* 19, 289–306 (1906).
6. Einstein, A. in *The Collected Papers of Albert Einstein Volume 1: The Early Years, 1879–1901* (eds Stachel, J. et al.) (Princeton Univ. Press, Princeton, 1987).
7. Stachel, J. *Fresnel's (Dragging) Coefficient as a Challenge to 19th Century Optics of Moving Bodies* (preprint 283, Max-Planck-Institut für Wissenschaftsgeschichte, Berlin 2004).
8. Einstein, A. in *The Collected Papers of Albert Einstein Volume 1: The Early Years, 1879–1901* (eds Stachel, J. et al.) 31–32 (Princeton Univ. Press, Princeton, 1993).
9. Kuhn, T. S. *Black-Body Theory and the Quantum Discontinuity, 1894–1912* Ch. 9 (Oxford Univ. Press, Oxford, 1978).

nicht ändert. Wir

$$L \left\{ \frac{1}{\sqrt{1 - \left(\frac{v}{V}\right)^2}} - 1 \right\}.$$

des Körpers in bezug auf (ξ, η, ζ) nimmt
 ... ab, und zwar um einen von den
 Körpers unabhängigen Betrag. Die Differenz
 ... ferner von der Geschwindigkeit ebenso ab wie
 ... Energie des Elektrons (l. c. § 10).
 ... Vernachlässigung von Größen vierter und höherer
 ... können wir setzen:

$$K_0 - K_1 = \frac{L}{V^2} \frac{v^2}{2}.$$

Aus dieser Gleichung folgt unmittelbar:

Gibt ein Körper die Energie L in Form von Strahlung ab, so verkleinert sich seine Masse um L/V^2 . Hierbei ist es offenbar unwesentlich, daß die dem Körper entzogene Energie gerade in Energie der Strahlung übergeht, so daß wir zu der allgemeineren Folgerung geführt werden:

Die Masse eines Körpers ist ein Maß für dessen Energieinhalt; ändert sich die Energie um L , so ändert sich die Masse in demselben Sinne um $L/9 \cdot 10^{20}$, wenn die Energie in Erg und die Masse in Gramm gemessen wird.

Es ist nicht ausgeschlossen, daß bei Körpern, deren Energieinhalt in hohem Maße veränderlich ist (z. B. bei den Radiumsalzen), eine Prüfung der Theorie gelingen wird.

Wenn die Theorie den Tatsachen entspricht, so überträgt die Strahlung Trägheit zwischen den emittierenden und absorbierenden Körpern.

Bern, September 1905.

(Eingegangen 27. September 1905.)

Einstein as icon

John D. Barrow

How Einstein became the personification of physics.

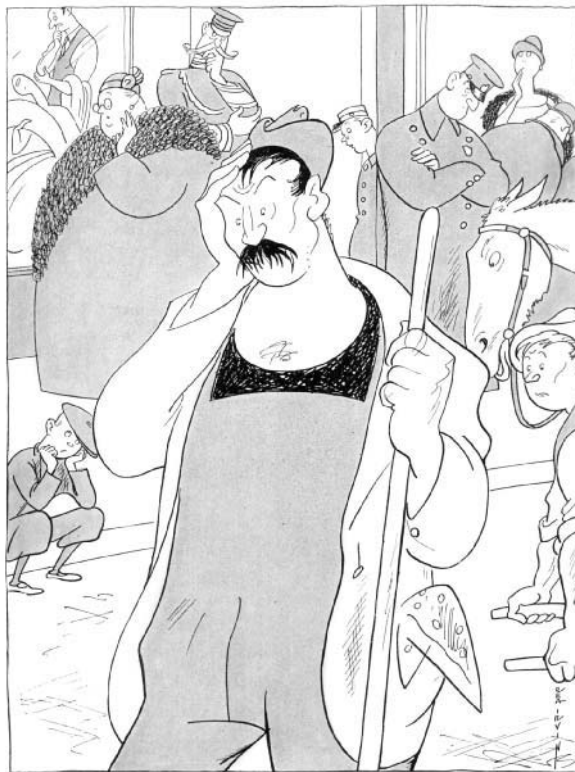
Scientific celebrity has a relativity all of its own. Some scientists are celebrated by their peers, some are treasured by their students, while others are lionized by the public at large. But very few are given the burden of being a celebrity to everyone, everywhere, all the time. Albert Einstein achieved that universality. Publicists must despair that he did so in the pre-television era, without an agent or the help of a public-relations company. He didn't even have a website. More striking still, I doubt that more than a handful of *Nature* readers have ever seen a colour photograph or a film of Einstein, and even fewer know what his voice sounded like. Yet his face has become an icon for wisdom, imagination, creativity and concentrated mental power; his brand name so pervasive a synonym for genius that even he once had to confess: "I am no Einstein."

There have always been scientific celebrities. Isaac Newton was the ornament of his age: people talked about him, made fun of him, and even devised newtonian models of government and morals. But Newton never became an icon. Instead, he remained severe and aloof. In the nineteenth century, Charles Darwin became, with Thomas Huxley's assistance, a reluctant public intellectual. Although he kept to the wings, his ideas were centre stage in long-running controversies about science and religion, the origins of humankind and our relationship to the hairier members of the animal kingdom. In retrospect, what is interesting about Darwin's work is that at some popular level it was too well known. People thought that they could understand what he was saying. Indeed, that was the problem — they understood too well.

Einstein restored faith in the unintelligibility of science. Everyone knew that Einstein had done something important in 1905 (and again in 1915) but almost nobody could tell you exactly what it was. When Einstein was interviewed for a Dutch newspaper in 1921, he attributed his mass appeal to the mystery of his work for the ordinary person: "Does it make a silly impression on me, here and yonder, about my theories of which they cannot understand a word? I think it is funny and also interesting to observe. I am sure that it is the mystery of non-understanding that appeals to them... it impresses them, it has the colour and the appeal of the mysterious."

Relativity was a fashionable notion. It promised to sweep away old absolutist notions and refurbish science with modern ideas. In art and literature too, revolutionary changes were doing away with old conventions and standards. All things were being made new. Einstein's relativity suited the mood. Nobody got very excited about Einstein's brownian motion or his photoelectric effect but relativity promised to turn the world inside out.

The fact that it was supposedly understood by a mere handful of people on the planet was a comfort and added cachet. You didn't need to try to understand and no one would think you stupid for not knowing. Especially appealing was the fact that Einstein's ideas about space and time used familiar words, such as mass and energy, but endowed these words with new and richer meanings. Everyone knew what



"People slowly accustomed themselves to the idea that the physical states of space itself were the final physical reality."
PROFESSOR ALBERT EINSTEIN

these words meant, but sentences, such as "everything is relative" or "moving clocks run slow" did not by themselves convey useful information. This irony was beautifully captured by Rea Irvin's cartoon for the *New Yorker Magazine* in 1929, which showed a New York street scene composed of a baffled road sweeper scratching his head surrounded by tradesmen and passing shoppers in various poses of puzzlement. The caption quotes Einstein: "People slowly accustomed themselves to the idea that the physical states of space itself were the final physical reality."

Einstein's celebrity began in 1919 when the British Eclipse expeditions confirmed his prediction of the bending of light by the Sun. This was headline news and in the years following the misery of the First World War it recreated a sense of order and hope about the Universe and human enquiry. Pictures of Einstein then were quite different. In 1905, he was a dapper young man working in a Swiss patent office. He was usually shown in a smart European-style suit, wing collar and bow tie, and would be instantly recognizable as an archetypal European professor.

After the Second World War, Einstein moved to Princeton where he became increasingly isolated from the main currents of development in theoretical physics because of his aversion to quantum theory. With no research group or school of students, he pursued a solitary search for an extension of the general theory of relativity. His aim was to unite relativity with the theory of electromagnetism and

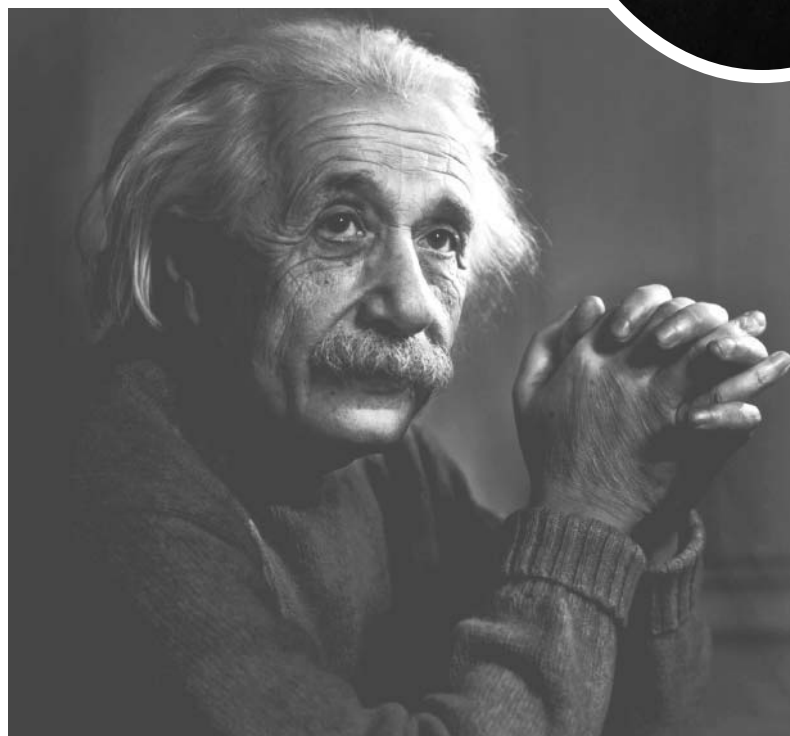


"Instinct says beer. Reason says Carlsberg."

produce a 'unified field theory' — the forerunner of today's 'theories of everything'. It is from this latter part of his life that his iconic legacy principally derives. The slightly bohemian looking, dishevelled old gentleman with the sad eyes — captured so beautifully by Yousuf Karsh's camera in 1950 — became the symbol of the power of the mind.

There are several things about Einstein's delayed celebrity that are interesting. First, Einstein's image in America contrasts sharply with that of the smart young man from Switzerland who made the acclaimed discoveries. The ubiquity of the older Einstein image meant that the concept of the scientific genius became associated with the image of an old fatherly figure.

The other great Einstein transformation, which occurred as the icon was emerging, is in some ways more interesting. This was in the style of his scientific work. In the brilliant papers of 1905, we see the hallmark of the young Einstein: penetrating ideas, motivated by crucial thought-experiments and finally



fashioned into perfect form by mathematics that is "as simple as possible but no simpler". Yet the development of the general theory of relativity introduced Einstein to the power of abstract mathematical formalisms, notably that of tensor calculus. A deep physical insight orchestrated the mathematics of general relativity, but in the years that followed the balance tipped the other way.

Einstein's search for a unified theory was characterized by a fascination with the abstract formalisms themselves. Rather than using physical arguments or thought-experiments, the strength and degrees of completeness of the formalisms themselves were used to sift their appropriateness as physical theories.

After his death in 1955, Einstein became the epitome of intelligence in general, not just of scientific genius. His image appeared on postage stamps and magazine covers. He became the subject of poetry, sculpture and painting. 'Einstein advertisements' implied that buyers were smart to buy the product, or likely to become smart if they did. Conversely, the simplicity and ease of use of a product was endorsed by the assurance that you did not need to be an Einstein to use it.

The most elaborate advertising campaign that exploited Einstein's image was the 1979 Pennaco Hosiery catalogue, intriguingly entitled "The Theory of Relativity — Fall-Winter 79". Women were exhorted to "Believe in the theory of relativity". "Fashion relates beautifully — making each part relative to the next — and thus creating fantastic total outfits. Thank you, Mr. Einstein, for the theory... and Happy 100th Birthday!"

Today, the Einstein industry is as strong as ever. I was surprised to find that I owned an Einstein suit, an Einstein toy, and inevitably an Einstein calculator. If the human race had to agree on an image to transmit to an extraterrestrial civilisation, I suspect that Einstein would be a part of it.

Most amazing of all is that — despite the hullabaloo and the inevitable cynicism about celebrity in our age, especially in response to media-created icons — Einstein's scientific legacy is greater than ever. His predictions about gravity have steadily been confirmed with just a few remaining beyond the reach of our experiments. His early scientific work was an unqualified success and his personal demeanour and response to fame an object lesson to all. This is why 2005 is the World Year of Physics — Einstein's Year.

John D. Barrow is at the Centre for Mathematical Sciences, University of Cambridge, Wilburforce Road, Cambridge CB3 0WA, UK. e-mail: j.d.barrow@damtp.cam.ac.uk

Brownian motion

Giorgio Parisi

**"I did not believe that it was possible to study the Brownian motion with such a precision."
From a letter from Albert Einstein to Jean Perrin (1909).**

On 30 April 1905, Einstein completed his doctoral thesis on osmotic pressure, in which he developed a statistical theory of liquid behaviour based around the existence of molecules. This work, together with his subsequent paper on 'brownian motion', constitutes one of the most important, but often overlooked, contributions that Einstein made to physics.

In the closing decades of the nineteenth century, theoretical physics was in a state of turmoil. The big outstanding questions of that time have been much discussed (including within this collection). Such questions culminated in relativity and quantum mechanics — theoretical developments in which Einstein's key role is being justly celebrated this year. But it should not be forgotten that Robert Brown's seemingly innocuous observations of the irregular motions of a suspension of pollen grains in water — now known as brownian motion — also heralded a revolution in physical thought.

Although the concepts of atoms and molecules are now universally accepted, this was not the case at the turn of the twentieth century. Ludwig Boltzmann's statistical interpretation of the laws of thermodynamics — a body of work deeply rooted in the ensemble dynamical motion of material atoms — had many adherents. But there were also many heavyweight dissenters (for a time including Max Planck), who did not accept that thermodynamics had its origins in the reversible motion of invisible hypothetical particles. And many distinguished physicists of the time (among them Wilhelm Roentgen) suspected that brownian motion indicated a clear failure of Boltzmann's formulation of the second law of thermodynamics.

It was in this context that Einstein's explanation for brownian motion made an initial impression. In particular, Einstein showed that the irregular motion of the suspended particles could be understood as arising from the random thermal agitation of the molecules in the surrounding liquid: these smaller entities act both as the driving force for the brownian fluctuations (through the impact of the liquid molecules on the larger particles), and as a means of damping these motions (through the viscosity experienced by the larger particles). This connection between displacement, $x(t)$, and the viscosity, η , can be expressed (in one dimension) as: $\langle x(t)^2 \rangle = RT t / (3N\pi a\eta)$, where R is the universal gas constant, N is Avogadro's number ($2R/3N$ is Boltzmann's constant k_B), T is the temperature and a is the radius of the suspended particles. This finding went beyond simply confirming the existence of atoms and molecules, and provided a new way of determining Avogadro's number. As Einstein himself remarked, the consequence of this relation is that one can see, directly through a microscope, a fraction of the thermal energy manifest as mechanical energy. By proving that a statistical mechanics description could explain quantitatively brownian motion, all doubts concerning Boltzmann's statistical interpretation of the thermodynamic laws suddenly faded.

But this was not the end of the story. Einstein's realization that the fluctuations responsible for the agitation of the suspended particles were the same as those responsible for the friction experienced by the particles in motion — the first

example of a 'fluctuation–dissipation' theorem — had far-reaching consequences for other systems in equilibrium.

Generalizing this idea to other systems was easy and it became a recurrent theme in Einstein's papers: in an equilibrated system under a small perturbation there is balance between a systematic force and a chaotic fluctuating force. This is the essence of the fluctuation–dissipation theorem at equilibrium that was later derived in full generality by Kubo.

This celebrated classical fluctuation–dissipation theorem can be proved microscopically, with the details of the proof depending on the dynamics of the system. The theorem is valid for all dynamical systems in which the Boltzmann distribution is reached at equilibrium. A further extension of the basic idea occurred as recently as the 1980s when it was realized that, in certain cases, the fluctuation–dissipation theorem was a consequence of a hidden supersymmetry and time translation invariance.

When we go out of equilibrium, the formal proof fails, and new properties and behaviour appear. In some systems, the departure from equilibrium is substantial and we must resort to different tools to analyse their behaviour. A good example is turbulence, where energy is continuously injected into the system, before cascading from one length scale to another.

But the situation is different for systems that are only slightly out of equilibrium. For example, imagine a system that cannot reach equilibrium because of high free-energy barriers (that may be of energetic or of entropic nature): this situation typically applies to disordered systems, such as spin glasses and structural glasses. Such a system will approach equilibrium slowly, by jumping from one metastable state to another, and it could remain slightly out of equilibrium forever if continually perturbed with a slowly changing external field.

In such systems we can expect a separation, by many orders of magnitude, between the microscopic time scale of the system (for example, that represented by the vibrations of individual atoms) and the macroscopic time needed to cross the barrier (for example, changes in the structure of the system itself). The system can then be considered to be essentially thermalized inside a metastable state, and so fluctuation–dissipation ideas can still be applied: the slowly changing overall state of the system is considered to be a small perturbation.

Although a formal analytical solution to such behaviour has yet to be found, it is encouraging that much of the off-equilibrium behaviour of disordered systems can be understood within a fluctuation–dissipation framework that in some respects constitutes a generalization of Boltzmann's equilibrium statistical mechanics. These ideas are explored in more detail for the specific case of glasses in this issue (page 222). ■

Giorgio Parisi is in the Department of Physics, Università di Roma La Sapienza, Piazzale Aldo Moro 5, Roma 00185, Italy.

FURTHER READING

Pais, A. *Subtle is the Lord...* (Oxford Univ. Press, 1982).

Kuhn, T. S. *Black Body Theory and the Quantum Discontinuity 1894–1911* (Oxford Univ. Press, 1978)

Mézard, M., Parisi, G. & Virasoro, M. A. *Spin Glass Theory and Beyond* (World Scientific, Singapore, 1987).

In and out of equilibrium

J. Kurchan

PMMH Ecole Supérieure de Physique et Chimie Industrielles, 10 rue Vauquelin, 75231 Paris CEDEX 05, France (e-mail: jorge@pmmh.espci.fr)

Albert Einstein's work on brownian motion showed how thermal equilibrium could be brought about by work exchanged through thermal fluctuations and viscous dissipation. Glasses are out-of-equilibrium systems in which this exchange happens at widely different timescales simultaneously. Theory then suggests the fascinating possibility that such behaviour may lead to a more general form of thermalization, in which the effective temperature shared by all components differs at each timescale.

Thermodynamics is a 'gift' of mechanics. Both newtonian and quantum dynamics in many-body systems lead to extremely complicated dynamical evolution. But both also have a feature that is exceptional amongst dynamical systems: in most cases, we can assume that all system configurations of a given energy will be visited with equal probability. When this assumption is valid, we say that the system is in equilibrium. We can then forget the evolution of the system in time, and describe its properties statistically — a procedure leading to thermodynamics. Disregarding the time-dependence involves a certain degree of technical simplification; but much more importantly, the statistical description brings about new, important concepts such as temperature and entropy. The price one pays is that certain quantities (such as diffusivity and viscosity) that do depend on the dynamical details cannot be studied with thermodynamics alone.

Einstein's papers on brownian motion showed a surprising fact for the first time: thermodynamics still contains dynamic information. A particle in suspension in a liquid is rocked by the irregularity of molecular collisions, but the viscous forces produced by the same molecules tend to dampen the motion. Energy is gained through the fluctuations (the irregular collisions) and lost through the dissipation (the viscous drag). The requirement that the particle be in equilibrium with the liquid then implies a balance between the diffusivity of the particle and the viscosity of the liquid — two purely dynamic quantities.

Contrast this situation with the case of glasses. Glasses are systems out of equilibrium, and a thermodynamic description of them is in principle not possible: strictly speaking, we are not even allowed to refer to 'the' temperature of a glass bottle. This is a major failure of thermodynamics.

But as I will discuss below, by taking into account heat exchange through fluctuations and dissipations between the constituents of a glass (with each process operating at different timescales), we are making progress towards a generalization of thermodynamics that could apply to this difficult problem.

Losing equilibrium

If cooled fast enough, liquids can be taken below their freezing temperature without crystallizing. When the temperature of a supercooled liquid is further lowered, the viscosity increases dramatically; but in contrast to the case of crystallization, this change hardly manifests itself in the microscopic structure. Rather, it becomes evident as a sharp slowing down of particle diffusion. The resulting sluggishness of molecular rearrangement makes equilibration of the system difficult. And as temperature is further lowered, there sooner or later comes a point when the liquid cannot keep pace with the changes in the thermal bath, and falls out of equilibrium: a glass is born.

If at any given point in time we keep the temperature fixed and wait, a glass will work its way slowly back to equilibrium — a process known as physical ageing. During ageing, a glass's viscosity gradually increases until it eventually reaches the equilibrium value at the new fixed temperature, but the time needed for this to happen can be astronomical. We are then inevitably led to ask whether there is a temperature below which no amount of waiting will lead to equilibration in a liquid state, and ageing essentially goes on forever — or until crystallization intervenes. There seems to be no answer yet in sight for this hard, inevitably fascinating (although in practice rather irrelevant), question.

Another way of perturbing a supercooled liquid is to force it to flow. When this happens, velocity gradients develop in the liquid, which then responds by lowering its viscosity — a phenomenon present (and welcome) when paint is spread and when blood flows in capillaries. What is important here is that this 'shear-thinning' phenomenon entails a strong deviation from equilibrium. A rough way to understand this is the following: because ageing involves a viscosity increase, a forced decrease in viscosity can be thought of as having a rejuvenating effect. For a glass, this means being forced further away from equilibrium. In a sense that will be made more precise below, a deeply supercooled liquid 'realises' that it is almost a glass as soon as we try to make it flow.

In practice, we know that a system is not in equilibrium if some quantity (typically energy or density) is not constant in time, or, more generally, if thermal currents inside the system have not yet died away. But as we do not expect any observable of an old piece of glass to be wildly evolving, or that strong thermal currents are still flowing within its bulk, we may wonder why glasses are usually described as systems that are far from equilibrium. A clarifying approach to this question, which takes us back to Einstein's brownian motion papers (and to some recent developments in glass theory), is to think of equilibrium as a situation in which every conceivable type of thermometer coupled to a part of the system reads the same temperature. We declare a system far from equilibrium if two different thermometers indicate substantially different temperatures in the same region of space. This is exactly what happens in glasses.

Because physicists are much more at home when they are able to apply the well-developed equilibrium techniques, the more traditional theoretical approach to glasses has been to consider these systems in an ideal limiting situation, in which they have aged at each temperature for sufficiently long to equilibrate — essentially giving up the description of a main aspect of glassiness. But it should be borne in mind that, the lower the temperature, the more unrealistically long this waiting time is; and that, even if equilibration is achieved, it will be destroyed by shear-thinning as soon as the liquid is made to flow.

Two timescales

If we follow the motion of a particle in a regular crystalline solid, we see it performing fast vibrations around its assigned position in the lattice, and only very rarely swapping sites with a neighbour. In contrast, a particle in a normal liquid wanders away from any given position at a fairly uniform rate. As we take the liquid deeper into the supercooled region, the motion starts to change in several ways. Most strikingly, it splits into a fast vibrational ('cage' or 'rattling') motion of amplitude comparable with the particle size, superimposed on a wandering motion ('structural relaxation') that is on average much slower. As the temperature is further decreased, the rattling continues in much the same way, but the time taken for a particle to wander increases dramatically: the timescales for 'cage' and 'structural' motion become more distinct. (The structural rearrangements become also more spatially correlated: there are small epidemics of mobility continuously spreading and dying out in the sample.) This behaviour is reflected in all observables: local quantities — such as density, magnetization or electric polarization — also fluctuate in two well-separated timescales (Fig. 1).

After a sudden reduction in temperature, the lengthening of the 'slow' timescale is not immediate: the system ages gradually into a situation of slower structural rearrangements, and this is accompanied by the aforementioned increase in viscosity. Conversely, when a supercooled liquid undergoes shear-thinning, the decrease of the viscosity is the result of a shortening of the structural rearrangement timescale, while the fast motion is practically unaffected. All this strongly suggests that if we are trying to understand the elements that are specific to glassiness, especially its out-of-equilibrium nature, it is on the slow part of the motion that we have to concentrate, the fast motion being almost the same as in an equilibrated system. But we also see that the very same actors — the particles — are performing both fast and slow motion simultaneously: is it possible that the same particle is both in and out of equilibrium?

Our way to probe equilibration is by means of thermometers, so to make the discussion more quantitative, we have to think about how temperature is measured within any system.

Thermometers and the fluctuation–dissipation theorem

Consider the simplest kind of thermometer: a harmonic oscillator of frequency Ω , such as a mass attached to a spring (or brownian particles in the springless limit $\Omega = 0$). To measure the temperature (T) of a system, we couple the oscillator to any of its components. In equilibrium, equipartition of energy tells us that, on average, the oscillator's kinetic energy must be $1/2 k_B T$ (where k_B is Boltzmann's constant), so it suffices to measure its mean value over a sufficiently long time to deduce T . Equilibrium thermodynamics guarantees that T will be independent of the oscillator's frequency Ω and also of the details of its coupling to the system.

There are situations in which we wish to couple the thermometer to a microscopic physical quantity X (for example, a local magnetization, electric polarization or density fluctuation) and ensure that the thermometer disturbs the evolution of this quantity as little as possible. At the level of thought experiment, one way to do this is to imagine that we have M copies of the system $A_1, A_2, \dots, A_M$ that have gone through the same experimental procedure, and the thermometer couples to the observable X corresponding to each copy of the system simultaneously (Fig. 2): if M is large, then the coupling to each individual copy will be very small. We may assume that X fluctuates around zero, and has on average correlation $\langle X(t)X(t') \rangle = C_X(t - t')$, or $C_X(\omega)$ in Fourier space. The fluctuations of X act as a noise that pumps energy into the thermometer, just like the irregularities in molecular collisions push a brownian particle.

However, this is not the end of the story. Each system A_i feels the presence of the thermometer as a small perturbation, and this on average modifies the X s. How much exactly X responds on average at time t to a small perturbation at time t' is measured by the so-called response function $R_X(t - t')$ (or $R_X(\omega)$ in Fourier space). This modification in

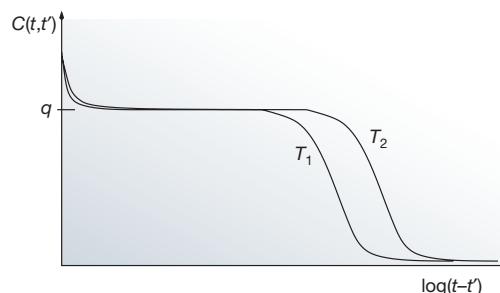


Figure 1 The two-time (t, t') correlation $C(t, t')$ of a typical physical quantity in terms of $\log(t - t')$ for two temperatures $T_1 > T_2$. The initial step ($C > q$) corresponds to the fast vibrational motion, and the second ($C < q$) reflects the slower structural relaxation (see text). The timescale for the latter is not only very sensitive to temperature, but also to shear rate — or, if the system is ageing, to the time elapsed since it fell out of equilibrium.

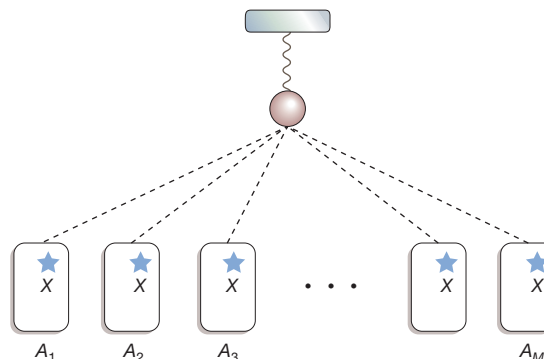


Figure 2 M copies of a system ($A_1, A_2, \dots, A_M$). The thermometer (an oscillator) is coupled to the physical quantity X in each copy of the system.

the X s feeds back into the thermometer, and tends to decrease its energy: this is the 'dissipation', which in the case of the agitated brownian particle represents the role of the liquid's viscosity. In summary, the thermometer feels the 'bare' noise of the observable — the 'fluctuations' (measured by $C_X(\omega)$) — and the echo of its own action on the system (proportional to $R_X(\omega)$). These two contrary effects must be such that they give the right energy as predicted by equipartition, for every conceivable thermometer and observable. A short calculation shows that this can only happen if correlations and responses associated with any observable are proportional: $T = [\omega C_X(\omega)]/[2\text{Im}R_X(\omega)]$. This is the fluctuation–dissipation theorem¹, a generalization of the relation found by Einstein between the diffusivity of brownian particles and the viscosity of the liquid.

Multi-thermalization

As we have just seen, a thermometer that selectively responds to a frequency Ω when coupled to a physical quantity X will indicate a temperature corresponding to the fluctuation–dissipation ratio: $T(\Omega, X) = [\Omega C_X(\Omega)]/[2\text{Im}R_X(\Omega)]$. It is then natural to refer to the quantity $T(\Omega, X)$ as the 'effective temperature' of X at timescale $1/\Omega$. In a system in equilibrium, all the $T(\Omega, X)$ are equal to the ambient temperature T .

In a situation of well-separated timescales, a first result — which can indeed be proved quite generally — is that a thermometer that

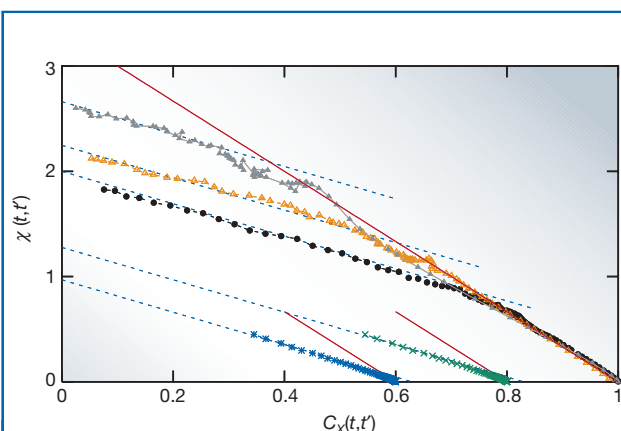


Figure 3 Fluctuation–dissipation plot for a sheared supercooled liquid, taken from ref. 5. The results are from a simulation of a mixture of two types of particle (see text); $\chi(t, t')$ is the integrated response, and $C_X(t, t')$ is the correlation function. Shown in black (filled circles), orange (open triangles) and grey (filled triangles) are the correlation of Fourier components of the structure factor for wave vector $k = 7.47$, $k = 11.22$ and $k = 14.48$, respectively. The crosses in green and blue are plots of the negative of the mobility versus the self-diffusion for each type of particle (the origin is shifted for ease of comparison). The solid lines correspond to the equilibrium value $-1/T$, while the dashed lines are all parallel with the same gradient $\equiv -1/T_{\text{eff}}$. The two-temperature model seems to fit the data well. (The fast timescale is not visible in the self-diffusion, as it implies distances too small to be visible on this plot.)

takes heat from the system only through the higher-frequency fluctuations (the ones that are least affected by departures from equilibrium) will indicate the ambient temperature T when coupled to any observable. The heat sensors in our skin are an example, and this explains why glasses feel no hotter to the touch than substances in equilibrium.

The out-of-equilibrium dynamics of glassy systems has been solved within a set of approximations that we shall describe in some detail below^{2–4}. A striking outcome is that fluctuations at long timescales comparable to that of structural relaxation are at an effective temperature T_{eff} that is higher than the bath temperature T — the same one for all physical quantities. In an ageing system, T_{eff} is of the order of (but not equal to) the temperature that the system had at the point it fell out of equilibrium, and gradually decreases with time to T if and when the system re-equilibrates. In a flowing supercooled liquid, the effective temperature becomes different as soon as shear-thinning appears. This two-temperature (T, T_{eff}) description is only meaningful to the extent that the timescales of fast ‘rattling’ motion and slow ‘structural rearrangements’ remain well separated. (One can of course conceive of cases with more than two well-separated timescales, each with its own effective temperature.) We are still very far from being able to validate this description exactly in a realistic model of a glass, but numerical simulation results for realistic glass models, and to some extent the results of experiments, are encouraging.

A way to interpret the fluctuation–dissipation data at a glance is the following³: for a physical quantity X , we measure the correlation function $C_X(t, t')$ and the integrated response:

$$\chi(t, t') = \int_{t'}^t d\tau R_X(t, \tau)$$

In an equilibrium system, the fluctuation–dissipation theorem states that a parametric plot of $\chi(t, t')$ against $C_X(t, t')$ yields a straight line with slope $-1/T$. If an out-of-equilibrium system has a two-step correlation, as in Fig. 1, and temperatures T and T_{eff} associated with the fast and slow timescales, respectively, the $\chi(t, t')$ versus $C_X(t, t')$ plot will be composed of two straight segments of slopes $-1/T$ and $-1/T_{\text{eff}}$ corresponding to short and long times. Clearly, if there is a

single T_{eff} for all physical quantities, the segments associated with the long-time decay of all quantities must be parallel. Figure 3 shows the results of a simulation of a mixture of two types of particle interacting via a Lennard–Jones potential, subjected to shear⁵. Similar results have been obtained for granular matter⁶. The results are quite reassuring.

Experimental results are also encouraging but currently contain much less detail^{7–9}, mainly owing to difficulties in accessing the response and correlation functions of many observables. Such behaviour is sometimes further obscured by the presence of a strong, intermittent, ‘crackling’ noise, an interesting and as yet not well-understood phenomenon, which seems to be operating at an entirely different length scale — and is probably beyond the scope of microscopic glass theories⁸.

At the phenomenological level, the notion that glasses can be described by two temperatures — the ambient temperature and a higher ‘fictive’ temperature — has been around for a long time¹⁰. It seems natural to think that the theoretical developments correspond by and large to that basic idea. It is difficult, however, to make a direct comparison between a theoretical framework that imposes relations and constraints, and a phenomenological approach that has been optimized to fit the experimental data. For example, to be meaningful as temperatures, the effective temperatures have to coincide at each timescale for all observables, a restriction that has been largely abandoned in the ‘fictive temperature’ phenomenological approaches.

‘Mean-bath’ theory

It is possible to understand the approximations that yield the effective temperature without going into too much technical detail. The argument also suggests why the same ideas could apply more generally.

In Fig. 2, the systems $A_1, A_2, \dots, A_M$ are assumed to be equilibrated at temperature T , so all the correlations and their associated response functions obey fluctuation–dissipation relations. I have argued above that these relations imply that thermalization is inherited by the thermometer, which in turn implies that all observables belonging to the thermometer itself should also satisfy fluctuation–dissipation relations. We could now couple the thermometer (or a set of such thermometers) to yet another system, and the inheritance of thermalization at temperature T will be carried one generation further. All this is very natural, but we can now ask a less obvious question: what would happen if the systems $A_1, A_2, \dots, A_M$ had two timescales with two different effective temperatures? Would this property be inherited in the same way? The answer turns out to be ‘yes’, provided that the two timescales are widely separated.

Consider now an approximation in which we break the system into components, each of which is assumed to be thermally coupled to the fluctuations of the rest (like the thermometer in Fig. 2), so that a particular component ‘feels’ the others as a thermal bath. The argument becomes self-consistent when one assumes that all components behave on average in the same way². What we are therefore proposing is just like the familiar mean-field approximation, except that here the subsystems only interact via their fluctuations. What is transmitted is not the average value of the observables, but rather the time-dependent fluctuations around these averages: there is a ‘mean thermal bath’ rather than a ‘mean-field’.

It is now clear that equilibrium is a possibility. Placed in contact with a mean thermal bath that is in equilibrium, a particular component will equilibrate, which in turn justifies the assumption for all the other components. However, if the system is out of equilibrium, either because it is being driven or because it is ageing, and has widely separated timescales, a self-consistent solution with two temperatures is also possible as, by virtue of the ‘hereditary’ property described above, a component in contact with a bath having two temperatures will also develop two temperatures.

The theoretical situation with this ‘multi-thermalization’ idea is quite similar to that of ferromagnetism a century ago, when a mean-field approximation was developed that showed how spontaneous magnetization is possible. In that case, each spin is assumed

to interact with all the rest: if the ‘other’ spins are assumed to have positive magnetization, they force positive magnetization on the selected spin, in turn justifying the initial choice of magnetization for the rest. The lesson that emerged was that a symmetry of the system can be spontaneously broken — but it took many years to show that this could also happen in more realistic lattice models of the spin systems. Similarly, we now know that multi-thermalization schemes are realizable within the mean thermal bath approximations described above, but we have no analytic proof that it happens in an exact solution of a realistic model. The numerical checks are encouraging, but will never tell us whether the multi-thermalization holds strictly, or only as an approximation.

Conclusions

The idea that glasses can be described by two temperatures has a long history. Rethinking the whole question of heat exchange, in a manner close to Einstein’s in his work on brownian motion, has made possible a theoretical basis for these old phenomenological ideas. The fluctuation–dissipation results suggest that, in certain cases, there can be an underlying generalization of statistical mechanics with two or more temperatures. Indeed such an approach was proposed for granular matter some time ago^{6,11}.

In general, what thermodynamics offers is not so much the technical advantage of being able to disregard time — after all, this only reduces the dimensionality of our world from four to three — but rather a framework with powerful notions such as temperature, entropy and equilibration, together with inequalities and arrows of

time. Physicists feel so lost without them, that inevitably they talk about ‘hot’, ‘cold’, ‘entropy’ and ‘thermalization’ even in situations that are far from equilibrium, despite being aware that these terms are not really defined in those cases. Any generalization of thermodynamics that applies to glasses (which represent approximately half the solids that surround us) will be most welcome. □

doi:10.1038/nature03278

1. Kubo, R., Toda, M. & Hashitsume, N. *Statistical Physics II: Nonequilibrium Statistical Mechanics* (Springer, New York, 1991).
2. Sompolinsky, H. & Zippelius, A. Relaxational dynamics of the Edwards-Anderson model and the mean-field theory of spin-glasses. *Phys. Rev. B* **25**, 6860–6875 (1982).
3. Cugliandolo, L. F. & Kurchan, J. On the out-of-equilibrium relaxation of the Sherrington-Kirkpatrick model. *J. Phys. A* **27**, 5749–5772 (1994).
4. Cugliandolo, L. F., Kurchan, J. & Peliti, L. Energy flow, partial equilibration, and effective temperatures in systems with slow dynamics. *Phys. Rev. E* **55**, 3898–3914 (1997).
5. Berthier, L. & Barrat, J.-L. Nonequilibrium dynamics and fluctuation–dissipation in a sheared fluid. *J. Chem. Phys.* **116**, 6228–6242 (2002).
6. Makse, H. A. & Kurchan, J. Testing the thermodynamic approach to granular matter with a numerical model of a decisive experiment. *Nature* **415**, 614–617 (2002).
7. Israeloff, N. E. & Grigera, T. S. Low-frequency dielectric fluctuations near the glass transition. *Europhys. Lett.* **43**, 308–313 (1998).
8. Buisson, L., Ciliberto, S. & Garcimartin, A. Intermittent origin of the large violations of the fluctuation–dissipation relations in an aging polymer glass. *Europhys. Lett.* **63**, 603–609 (2003).
9. Herisson, D. & Ocio, M. Fluctuation–dissipation ratio of a spin glass in the aging regime. *Phys. Rev. Lett.* **88**, 257202 (2002).
10. Tool, A. Q. Relation between inelastic deformability and thermal expansion of glass in its annealing range. *J. Am. Ceram. Soc.* **29**, 240–253 (1946).
11. Edwards, S. F. in *Granular Matter: An Interdisciplinary Approach* (ed. Mehta, A.) 121–140 (Springer, New York, 1994).

Competing interests statement The author declares that he has no competing financial interests.

Quantum criticality

Piers Coleman¹ & Andrew J. Schofield²

¹Center for Materials Theory, Rutgers University, Piscataway, New Jersey 08854-8019, USA (e-mail: coleman@physics.rutgers.edu)

²School of Physics & Astronomy, University of Birmingham, Edgbaston, Birmingham B15 2TT, UK (e-mail: ajs@bham.ac.uk)

As we mark the centenary of Albert Einstein's seminal contribution to both quantum mechanics and special relativity, we approach another anniversary — that of Einstein's foundation of the quantum theory of solids. But 100 years on, the same experimental measurement that puzzled Einstein and his contemporaries is forcing us to question our understanding of how quantum matter transforms at ultra-low temperatures.

At the turn of the twentieth century a crisis less grandiose, yet just as profound as Michelson and Morley's failure to determine the Earth's motion through the ether, was troubling the founders of modern physics¹. The puzzle stemmed, not from astrophysical observations, but from the innocent question about the amount of energy required to raise the temperature of diamond; that is, its specific heat. The trouble was that much less energy was needed than expected. Almost a hundred years before, Pierre-Louis Dulong and Alexis-Thérèse Petit had observed that no matter what a solid is made of, its specific heat per molecule is roughly the same. In 1876 this was put on an apparently solid theoretical foundation by Ludwig Boltzmann. By applying his statistical mechanics to the atoms in a solid he was able to compute the specific heat per atom in precise agreement with Dulong and Petit's early observations. But by the last years of the nineteenth century this triumph of statistical physics was looking increasingly hollow. Not only did some materials, such as diamond, show too small a specific heat but the advent of cryogenic techniques showed specific heats to be strongly temperature dependent at low temperatures, in contrast to Boltzmann's theory. Not for the last time would low-temperature physics point to the need for a new understanding.

Attempts to resolve the clash between theory and experiment were espoused by three great physicists of the nineteenth century: Boltzmann, Lord Kelvin and Lord Rayleigh. Boltzmann suggested that the way atoms behaved when confined within a solid might not be as straightforward as he had assumed. Lord Kelvin, on the other hand, looked to the mathematics of Boltzmann's derivation, convinced that the error lay there. It was Lord Rayleigh who was prepared to say that both the experiment and the theory were correct and that there was a true crisis which could only be resolved by fresh insight. The next step to solving the problem was made by Einstein.

Einstein and the puzzle of specific heats

Einstein took his inspiration from an unlikely source — the light from stars. In 1900 Max Planck had developed a formula that could relate the colour of stars to their temperature — the first step in what would become quantum theory. Planck's theory had led Einstein to conclude that light is made up of discrete quanta or 'photons'. In 1906 Einstein applied Planck's formula to the vibrations inside matter which, he reasoned, must also be made up of quanta — tiny wave packets of sound that we now call 'phonons'. At high temperatures, Einstein's theory reverted to that of Boltzmann, but once the temperature dropped below that required to excite phonons, the heat content dropped dramatically. In his paper of November 1906 Einstein published one of his few fits to data — comparing his theory to the specific heat

of diamond — to demonstrate his solution of a seventy-year-old puzzle. A new discipline of the quantum theory of solids was born.

Einstein's 1906 theory was necessarily incomplete but it contained the seeds of an impending revolution. A first-principles derivation of Einstein's simple theory had to wait until 1924 when it became the first problem to be solved by Werner Heisenberg using the new 'quantum mechanics'. By 1930 the approach pioneered by Einstein had essentially wrapped up the specific heat capacity problem.

Quantum criticality and the return of the puzzle

A century later, the very same measurement of the specific heat of solids points to a new clash with our theories of matter. The 'bad actors' of the twenty-first century are man-made crystals at the forefront of modern materials science. Physics has traditionally focused on stable phases of matter, such as those shown by superconductors, magnets or ferro-electrics, but with modern materials it is possible to study unstable quantum phases of matter. Puzzling new behaviour develops around the precarious point of instability between two stable phases of matter: the 'quantum critical point'. The unique properties that develop in quantum critical matter have become a major focus of research in the past ten years. Moreover, the discrepancies that are unfolding between established theory and experiment leave us in a state of affairs that is curiously similar to that which existed with diamond a hundred years ago.

As the materials of our lives become more sophisticated — from complex plastics to advanced metal ceramics — one might think that understanding their properties would require knowing, in ever greater detail, the intricate complexities of electron and atom motion. Alas, such a task is impossible; even the most advanced computers are unable to cope with more than a few dozen electrons at once. Fortunately however, as Boltzmann foresaw when considering diamond, collections of particles can behave differently from isolated ones, and although their individual motions are complex, their collective properties acquire qualitatively new forms of simplicity. The appearance of magnetism, the development of zero-viscosity states in so-called superfluids, and the emergence of zero resistivity in a superconducting metal, are all examples of new, yet simple, behaviour that arise from the collective motion of electrons or atoms in complex matter. These are all examples of stable phases of matter. Collective behaviour also becomes important at the unstable interface between two stable phases of matter: the point of transformation is called a 'phase transition'.

Phase transitions are ubiquitous: from the crystallization of water into snowflakes, the alignment of electron spins inside a ferromagnet, the emergence of superconductivity in a cooled metal to the very formation of space-time in the early Universe; all involve phase transitions. Despite their

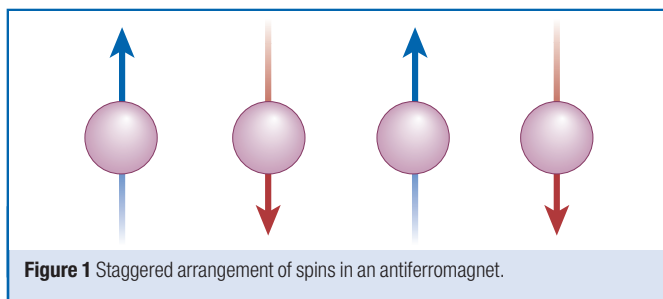


Figure 1 Staggered arrangement of spins in an antiferromagnet.

diversity, different phase transitions often share many fundamental characteristics. The specific heat when water turns to steam at a critical pressure has exactly the same power-law dependence on temperature as that of iron when it is demagnetized by having its temperature raised. Understanding this universal behaviour, known as ‘critical phenomena’, was a triumph of twentieth century physics². One of the key discoveries was that the imminent arrival of order at such continuous phase transitions is signalled by the formation of short-lived droplets of nascent order that grow as the system is tuned to the critical point. At the critical point, the material is spanned by droplets of all sizes.

The melting of ice is, like most phase transitions, caused by the increase in random thermal motion of the molecules which occurs as the temperature is raised. The ordered arrangement of atoms that exists in the solid cannot be sustained beyond a certain temperature and the crystal melts. Yet research into condensed matter over the past decade has revealed a new kind of phase transition that is driven, not by thermal motion, but by the quantum fluctuations associated with Heisenberg’s uncertainty principle. These quantum fluctuations are called ‘zero-point motion’. According to the uncertainty principle, the more certain a particle’s position, the more uncertain is its velocity. Thus, even when random, thermal motion ceases at the absolute zero temperature, atoms and molecules cannot be at rest because this would simultaneously fix their position and velocity. Instead they adopt a state of constant agitation. Like thermal motion, if zero-point motion becomes too wild, it can melt order, but in this case the melting takes place at absolute zero. Such a quantum phase transition³ takes place in solid helium, which is so fragile that it requires a pressure to stabilize its crystal lattice even at absolute zero. When the pressure is released, zero-point motions melt the crystal.

The best studied examples of quantum phase transitions involve magnetism in metals. Electrons have a magnetic direction or spin, which when aligned in a regular fashion makes a material magnetic. Iron magnetizes when all the spins inside align in parallel, but in other materials the spins form a staggered, alternating, or antiferromagnetic, arrangement (Fig. 1). These more fragile types of order are more susceptible to melting by zero-point fluctuations. Almost three decades ago theoretical physicist John Hertz, now at Nordita, made the first study of how quantum mechanics would affect phase transitions⁴. Hertz was fascinated by the question of how critical phenomena might be altered by quantum mechanics. Applying quantum mechanics to phase transitions turns out to be very like Einstein’s relativistic unification of space-time. In Hertz’s theory, quantum mechanics appears by including a time dimension to the droplets of nascent order. Normally this produces no additional effect, but Hertz reasoned that if a phase transition took place at absolute zero, then the droplets of order that foreshadow the transition would become quantum-mechanical rather than classical. At a zero-temperature phase transition, he reasoned, these quantum droplets would grow to dominate the entire material, changing its properties in measurable ways — and most affected would be the electrons (Fig. 2). Such ‘quantum critical matter’ offers the real prospect of new classes of universal electronic behaviour developing independently of the detailed material behaviour, once the material is driven close to a quantum critical point⁵.

The electrons that carry current in a metal are similar to photons of light: they are quantum waves whose wavelength decreases as their momentum increases. Unlike photons however, electrons obey the ‘Pauli exclusion principle’: no two electrons can share the same wavelength (or momentum). To minimize their energy, each electron must occupy the momentum state with lowest energy that is not already filled. When all electrons have done this, there is then a sharp demarcation (called the Fermi surface) separating the highest-energy occupied and lowest-energy unoccupied momentum states. This ordering imposes severe constraints: like apples packed into a barrel, only those near the top can be easily rearranged. One consequence is seen in the specific heat: only electrons near the Fermi surface can absorb thermal energy and find an unoccupied momentum state to move to. This means that the specific heat of metals is tiny, but that it grows linearly with temperature. The coefficient of linearity, called the Sommerfeld coefficient, is a rough measure of the ‘effective’ mass of the highest-momentum electrons inside the metal. When the effective mass is compared to that of an isolated electron it is usually found to be somewhat larger. This is because of the forces between electrons, which mean that as an electron moves around it has to push neighbouring electrons aside.

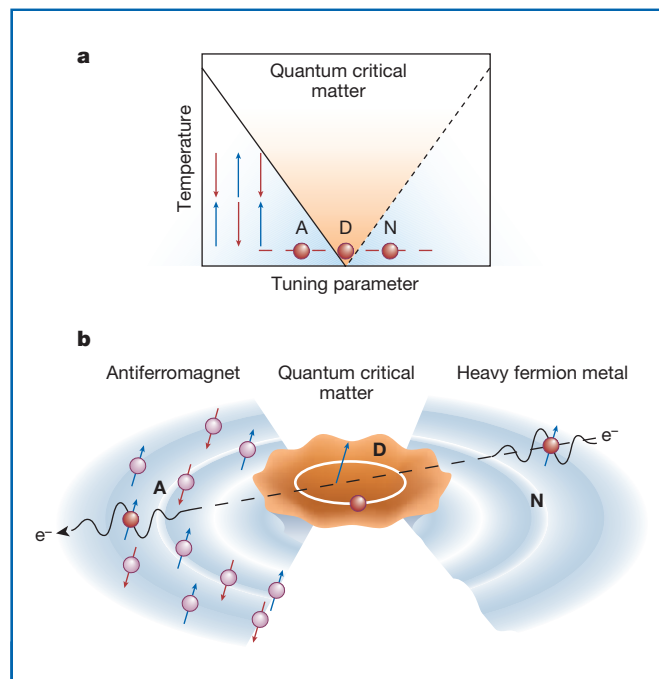


Figure 2 Schematic illustration of a quantum critical point showing the phase diagram (a) and the growth of droplets of quantum critical matter near the quantum critical point (b). **a**, Schematic phase diagram near a quantum critical point. Quantum critical points distort the fabric of the phase diagram creating a ‘V-shaped’ phase of quantum critical matter fanning out to finite temperatures from the quantum critical point. **b**, As matter is tuned to quantum criticality, ever-larger droplets of nascent order develop. On length-scales greater than these droplets, electrons propagate as waves. Inside the droplet, the intense fluctuations radically modify the motion of the electron, and may lead to it breaking up into its constituent spin and charge components. Physics inside the V-shaped region of the phase diagram (a) probes the interior of the quantum critical points (D), whereas the physics in the normal metal (N) or antiferromagnet (A) reflects their exterior. If, as we suspect, quantum critical matter is universal, then no information about the microscopic nature of the material penetrates into the droplets. Making an analogy with a black hole, the passage from non-critical, to critical quantum matter involves crossing a ‘material event horizon’. Experiments that tune a material from the normal metal past a quantum critical point force electrons through the ‘horizon’ in the phase diagram, into the interior of the quantum critical matter, from which they ultimately re-emerge through a second horizon on the other side into a new universe of magnetically ordered matter.

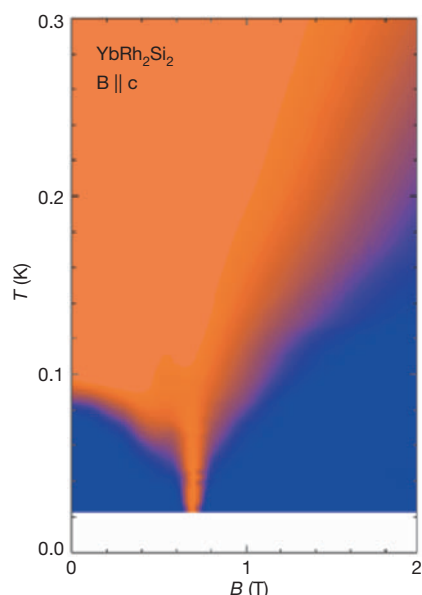


Figure 3 ‘Singularity in the phase diagram’ illustrated by data taken from the material YbRh_2Si_2 where an applied magnetic field tunes the material to a quantum critical point²⁵. Blue regions indicate normal metallic behaviour. Orange regions indicate anomalous metallic behaviour with linear resistivity. The singular quantum critical point at absolute zero produces a wide region of unusual metallic behaviour at finite temperatures.

But much more dramatic changes occur near a quantum phase transition. Ten years ago, Hilbert von Lohneyson’s group in Karlsruhe, Germany, decided to measure the specific heat of a quantum critical metal⁶. They chose CeCu_6 , which can be tuned to a magnetic — actually, an antiferromagnetic — quantum critical point by adding small amounts of gold. As they added gold to the metal, they found that the Sommerfeld ratio of the metal kept getting bigger — as if the approach to criticality led to a metal where the electrons were getting heavier and heavier. But at the quantum critical concentration, the Sommerfeld coefficient never settled down to a constant. It just kept on rising as they lowered the temperature, as if the mass of the electrons on the Fermi surface was becoming infinite and the energy of the electrons vanishing.

The Karlsruhe group found another disturbing property. In normal metals, the electrical resistance due to electrons scattering off one another grows as the square of the temperature, but in this system, at the quantum critical point, the resistivity is linear with temperature. The constancy of the Sommerfeld coefficient and the quadratic temperature dependence of the resistivity constitute two absolutely stalwart signatures of normal metals. Their breakdown suggests that a quantum critical metal is a fundamentally new type of electron fluid.

Since the original Karlsruhe measurements, many new discrepancies have come to light⁷. Quantum critical points have now been found by pressure tuning⁸ and by applying magnetic fields⁹. We even have an apparent case of a line of quantum criticality instead of a single point¹⁰. All support the notion that the characteristic energy scale of the metal is driven to zero at a quantum critical point. Indeed, temperature itself seems to be the only energy scale that remains in quantum critical matter. In neutron scattering¹¹, for example, the rate at which electrons are scattered off the critical magnetic fluctuations seems to depend solely on the ratio of energy to temperature. Linear, or quasi-linear, resistivity is another sign of this phenomenon and, in the more dramatic cases, it can be followed over three decades in temperature.

One thing becoming increasingly clear is that although matter can never be cooled down to the quantum critical point at absolute zero,

drastic effects are felt long before this point is reached (Figs 2 and 3). It is this influence that elevates quantum criticality from an intellectual abstraction at absolute zero to a real-world phenomenon that can profoundly change finite-temperature material properties. The quantum critical point represents a kind of ‘black hole’ in the material phase diagram and this proves a useful analogy. Just as the black hole of cosmology distorts surrounding space-time, quantum critical points distort the fabric of the phase diagram (Fig. 2), creating a ‘V-shaped’ region of quantum critical matter fanning out to finite temperatures from the quantum critical point.

Resolving the quantum critical enigma

Some of the strange properties we see in quantum critical metals echo the predictions of Hertz and subsequent extensions of his theory¹², but this conventional wisdom is flouted in at least two important respects. First, the effects are much stronger and wide ranging than Hertz expected. Second, the particular case of antiferromagnetism should be by far the least dramatic: the alternating pattern of up and down magnetism should average out for almost all electrons, slowing down only the tiny fraction that can interfere constructively with this arrangement. In marked contrast, the experiments tell us that the effective mass of every electron on the Fermi surface is being increased to infinity — in essence, bringing every electron to a halt. Like Einstein a century before us, we have a thirty-year-old theory that cannot account for present experiments.

The current theoretical alternatives also in some sense parallel those of Einstein’s contemporaries, and there is great controversy. Some believe that the Hertz theory can indeed be saved if proper account is taken of the complexity of the material. The Karlsruhe team have advanced the idea that the underlying antiferromagnetic spin fluctuations at a quantum critical point collectively arrange themselves to become two-dimensional¹³. In this case, the motion of electrons between magnetic layers becomes highly turbulent, and it is possible to understand how they are driven to a halt in quantum critical matter. The only problem is that there is no obvious way for the spin fluctuations to become stubbornly two-dimensional in metals like CeCu_6 where the electrons are not confined to layers.

Others seek to explain the discrepancy as a failure of Hertz’s original approximations¹⁴. They point out that the distinction between the fluctuating magnetism and the electrons that appears in the Hertz theory is not so obvious when you remember that the magnetism itself also comes from the same electrons. Add to that the realities that materials contain disorder (atoms out of place, for example) and the assumptions of the mathematical derivations may not be quite so convincing.

Finally, there is the third possibility that, as Lord Raleigh advocated in the case of diamond, there is a true crisis and a new framework for quantum phase transitions is required. That would certainly be remarkable since the current theory is a beautiful synthesis of two very successful foundation stones of modern physics: the theory of temperature-driven phase transitions and quantum mechanics. Yet in the face of such puzzling experimental observations we are being driven towards such a possibility.

Several fascinating ideas have come to the forefront. One suggestion is that we are seeing a new kind of quantum criticality that turns classical criticality on its head. Classical criticality involves the global growth of static order in space: Hertz’s theory takes this global model and also includes growth along the quantum time dimension. An alternative proposal is that quantum critical matter involves the growth of droplets of order in time, but not space¹⁵. The debate between proponents of ‘global’ and this ‘local’ quantum critical theory is highly contentious. Another possibility is that electrons break up inside the highly collective environment of quantum critical matter. It could be that electrons inside quantum critical matter break up into their constituent spin and charge components¹⁶ — like atoms splitting up into ions in solution. This is related to ‘deconfined criticality’ which has been used to describe quantum critical points in magnetic,

two-dimensional insulators¹⁷. The only problem is that no one yet knows how to apply these ideas in detail to truly three-dimensional quantum critical metals.

The puzzle of quantum criticality is tantalizingly important for material science because, quite frequently, electrons appear to preempt the intense critical fluctuations of a quantum critical point by re-organizing themselves at the last minute into a new stable phase of matter. The divergence of the effective masses of the electrons and the collapse of electron kinetic energies all point to the formation of a highly degenerate state at the quantum critical point making them very susceptible to transformation into alternative stable electronic configurations.

For this reason, quantum criticality may be a highly effective catalyst for the formation of new stable types of material behaviour, providing an important new route for the design and discovery of new classes of material. One tendency is towards superconductivity^{18,19}. High-temperature superconductors — which currently become superconducting at liquid nitrogen temperatures and inspire the hope for one day achieving room-temperature superconductivity — exhibit a linear resistivity like metals, up to their melting temperature. Some believe that high-temperature superconductors and a whole host of other superconductors with quasi-linear resistances in the normal state, are driven by quantum criticality. Paradoxically, it is hard to prove this hypothesis because any attempt to remove the surrounding order will, at the same time, destroy the quantum criticality. We are faced by what has been described as a ‘quantum conundrum’⁴.

Many other unidentified phases also appear to develop around quantum critical points. In $\text{Sr}_3\text{Ru}_2\text{O}_7$ (ref. 20) and URu_2Si_2 (ref. 21), magnetic fields are used to tune the materials to a quantum critical point. These magnetic fields are too strong to allow the electrons to use superconductivity to avoid the critical singularity. Yet, rather than face the quantum critical point itself, the electrons find another way out. Abrupt changes in the resistivity and even in the material shape signal that the electrons have adopted a new type of ordered state. Is this a new form of magnetic order²² or a re-arrangement of the electrons’ flow²³ or perhaps both? All we know is that the electrons are still mobile but the precise rules of their new collective motion are still to be fully characterized and understood.

A hundred years ago, Einstein looked to the stars for inspiration to understand the properties of cold, stable, quantum matter. Today, the mysterious properties of critically unstable quantum matter not only suggest a new pathway to new material design, but they also raise our hopes for a new link between matter in the laboratory and matter in the cosmos. It has recently been proposed that ideas of quantum criticality may link up with quantum gravity at the surface of a black hole²⁴. Moreover, the break-up of electrons inside quantum critical matter and the emergence of new forces closely parallels challenges faced by the particle physicist, who seeks to understand how new forces develop between particles as they break up at high energies. Just as the inspiration of photons from stars led Einstein to solve the mystery of specific heat in stable quantum matter a hundred

years ago, some of the ideas now being considered for the properties of particles in the early Universe, such as gauge theory and supersymmetry, may be poised to reappear, right under our nose in the laboratory, as the solution to the specific heat problem in quantum critical matter. □

doi:10.1038/nature03279

- Pais, A. *Subtle is the Lord: the Science and the Life of Albert Einstein* Ch. 20, 389–401 (Oxford Univ. Press, Oxford, 1982).
- Domb, C. *The Critical Point: a Historical Introduction to the Modern Theory of Critical Phenomena* (Taylor & Francis, London, 1996).
- Sachdev, S. *Quantum Phase Transitions* (Cambridge Univ. Press, New York, 1999).
- Hertz, J. Quantum critical phenomena. *Phys. Rev. B* **14**, 1165–1184 (1976).
- Laughlin, R. B., Lonzarich, G. G., Monthoux, P. & Pines, D. The quantum criticality conundrum. *Adv. Phys.* **50**, 361–365. (2001).
- von Lohneysen, H. *et al.* Non-Fermi-liquid behavior in a heavy-fermion alloy at a magnetic instability. *Phys. Rev. Lett.* **72**, 3262–3265 (1994).
- Stewart, G. R. Non-Fermi-liquid behavior in d- and f-electron metals. *Rev. Mod. Phys.* **73**, 797–855 (2001).
- Julian, S. R. *et al.* The normal states of magnetic d and f transition metals. *J. Phys. Condens. Matt.* **8**, 9675–9688 (1996).
- Grigera, S. A. *et al.* Magnetic field tuned quantum criticality in the metallic ruthenate Sr_2RuO_6 . *Science* **294**, 329–332 (2001).
- Doiron-Leyraud, N. *et al.* Fermi liquid breakdown in the paramagnetic phase of a pure metal. *Nature* **425**, 595–599 (2003).
- Schröder, A. *et al.* Onset of antiferromagnetism in heavy-fermion metals. *Nature* **407**, 351–355 (2000).
- Millis, A. J. Effect of a non-zero temperature on quantum critical points in itinerant fermion systems. *Phys. Rev. B* **48**, 7183–7196 (1993).
- Rosch, A. Interplay of disorder and spin fluctuations in the resistivity near a quantum critical point. *Phys. Rev. Lett.* **82**, 4280–4283 (1999).
- Belitz, D., Kirkpatrick, T. R. & Rollhäuser, J. Breakdown of the perturbative renormalization group at certain quantum critical points. *Phys. Rev. Lett.* **93**, 155701/1–4 (2004).
- Si, Q., Rabello, S., Ingersent K. & Smith, J. L. Locally critical quantum phase transitions in strongly correlated metals. *Nature* **413**, 804–808 (2001).
- Coleman, P., Pépin, C., Qimiao Si & Ramazashvili, R. How do Fermi liquids get heavy and die? *J. Phys. Condens. Matt.* **13**, R723–R738 (2001).
- Senthil, T., Vishwanath, A., Balents, L., Sachdev, S. & Fisher, M. P. A. Deconfined quantum critical points. *Science* **303**, 1490–1494 (2004).
- Mathur, N. D. *et al.* Magnetically mediated superconductivity in heavy fermion compounds. *Nature* **394**, 39–43 (1998).
- Petrovic, C. *et al.* A new heavy-fermion superconductor CeIrIn_5 : a relative of the cuprates? *Europhys. Lett.* **53**, 354–359 (2001).
- Grigera, S. A. *et al.* Disorder-sensitive phase formation linked to metamagnetic quantum criticality. *Science* **306**, 1154–1157 (2004).
- Kim, K. H., Harrison, N., Jaime, M., Boebinger, G. S. & Mydosh, J. A. Magnetic-field-induced quantum critical point and competing order parameters in URu_2Si_2 . *Phys. Rev. Lett.* **91**, 256401/1–4 (2003).
- Amitsuka, H. *et al.* Hidden order and weak antiferromagnetism in URu_2Si_2 . *Physica B* **312–3**, 390–396 (2002).
- Chandra, P. *et al.* Hidden orbital order in URu_2Si_2 . *Nature* **417**, 831–834 (2002).
- Chapline, G. & Laughlin, R. B. in *Artificial Black Holes* (eds Novello, M. *et al.*) 179–198 (World Scientific, Singapore, 2002).
- Custers, J. *et al.* The break up of heavy electrons at a quantum critical point. *Nature* **424**, 524–527 (2003).

Acknowledgements We gratefully acknowledge discussions with P. Chandra, Z. Fisk, A. P. Mackenzie and D. Pines. P.C. is supported by the National Science Foundation. A.J.S. is supported by the Royal Society, the Leverhulme Trust and the EPSRC.

Competing interests statement The authors declare that they have no competing financial interests.

Happy centenary, photon

Anton Zeilinger^{1,2}, Gregor Weihs¹, Thomas Jennewein² & Markus Aspelmeyer¹

¹Institute for Experimental Physics, University of Vienna, Boltzmanngasse 5, A-1090 Vienna, Austria (e-mail: zeilinger-office@quantum.at)

²Institute for Quantum Optics and Quantum Information (IQOQI), Austrian Academy of Sciences, Boltzmanngasse 3, A-1090 Vienna, Austria

One hundred years ago Albert Einstein introduced the concept of the photon. Although in the early years after 1905 the evidence for the quantum nature of light was not compelling, modern experiments — especially those using photon pairs — have beautifully confirmed its corpuscular character. Research on the quantum properties of light (quantum optics) triggered the evolution of the whole field of quantum information processing, which now promises new technology, such as quantum cryptography and even quantum computers.

Of the papers written by Einstein in his *annus mirabilis* (1905), it was not the one where he introduced the special theory of relativity¹, but the one where he proposed the idea of quanta of light², later called photons³, that received the acclaim of the Nobel committee. This paper is often presented as if Einstein, having analysed the photoelectric effect, arrived at the idea of the photon. Yet, as is so often the case, the real story is much more interesting (see Box 1).

Since 1905, the photon has come a long way, considering that it was first regarded only to be a 'mathematical trick' or a concept without any deeper meaning (Box 1). But what exactly do we mean by a 'photon' today and what experimental evidence do we have to support the concept of the photon?

Single photons as particles and waves

A basic meaning of the term 'photon' is that radiation only exists in quantized energy packets. This contrasts with semiclassical radiation theories (see Box 2), which propose that matter is ruled by quantum physics, while the radiation field is classical.

One essential experiment that discriminates the quantum theory of light from a semiclassical one uses a stream of single photons incident on a beam splitter (Fig. 1). A semiclassical theory predicts that the two detectors in the output beams sometimes register in coincidence; according to this theory, the probability of registering a count is proportional to the square of the electric field. In contrast, full quantum theory predicts that the two detectors never register in coincidence. The quantum mechanically predicted statistics were experimentally confirmed by Clauser in 1974 (ref. 4), who used sources that emitted photons in pairs. Here, the registration of one of the two photons in a trigger detector indicates that a second, single photon is available for the experiment (Fig. 1). In Clauser's and in other early experiments the source was an atomic cascade where two photons are emitted, one after the other, within the lifetime of the intermediate state, which in general is very short.

Today, the source of choice for photon pair creation is the process of spontaneous parametric down-conversion (SPDC), the inverse process of frequency doubling. Both SPDC and frequency doubling are nonlinear optical processes. Whereas in frequency doubling two photons are converted into one photon of higher energy, in SPDC one photon from a pump laser beam is spontaneously converted into two photons, which emerge simultaneously⁵. Here also, registration of one of the two photons can serve as a trigger to indicate that the second photon has been generated. This results in single-photon states to a good approximation, because higher-order emission processes are negligible.

Very early on, Einstein criticized the new nature of randomness in quantum physics, most unforgettably by stating: "God does not play dice." In the light of this randomness, he also said that he would prefer to be an employee in a casino than a physicist. How would he comment on the later finding that one can construct random number generators on the basis of a single photon and a beam splitter⁶, as just described? Such a quantum random number generator could well be used in a casino because of the high quality of its random sequences.

Clauser's experiment⁴ contains the first demonstration of sub-poissonian photon counting statistics, which can only be understood within a quantum theory of light. Further experiments showed other purely quantum-based effects, such as the observation of photon antibunching in a resonance-fluorescence experiment⁷. These early experiments used beams of atoms as sources, where fluctuations in the atom number, and thus in the emission statistics, are unavoidable. Later, Walther's group in Munich realized such experiments using single atoms in traps⁸.

Single-photon interference

One of the most fascinating phenomena is quantum interference with individual photons. The interference pattern is observed by sending particles, one by one, through, say, a double slit assembly; many particles are then collected at the observation plane. In the simplest of such experiments, the light intensity can be dimmed down far enough that only one photon at a time is inside the apparatus. This was first demonstrated by Taylor⁹. In his experiment Taylor simply had a very dim light source together with a double slit assembly and a photo plate inside a box. But the results of such experiments can easily be understood semiclassically without having to assume the existence of photons; that is, without having to quantize the electromagnetic field as discussed above.

A single-photon interference experiment was performed by Grangier *et al.*¹⁰, who also used photon pairs emitted by atomic cascades. He and his colleagues employed a Mach-Zehnder interferometer to observe real single-photon interferences. Figure 2 shows the results of a single-photon double-slit experiment^{11,12} where the photon source was parametric down-conversion. The intensities are extremely low; nevertheless, the interference pattern accumulated photon by photon shows perfect interference fringes. Experiments of this kind clearly confirm that the quantum state is not just a statistical property of an ensemble of particles; indeed, it makes very precise predictions even for individual particles. Following Feynman¹³, the fact that the predictions of quantum mechanics hold for individual particles and not just for ensembles is best illustrated by the

Box 1

A heuristic concept

The way that Einstein arrives at the photon concept in his seminal paper “Über einen die Erzeugung und Verwandlung des Lichtes betreffenden heuristischen Gesichtspunkt” (“On a heuristic aspect concerning the production and transformation of light”)² is, contrary to widespread belief, not through the photoelectric effect. Instead, Einstein compares the entropy of an ideal gas filling a given volume with the entropy of radiation filling a cavity. The logarithmic dependence on the volume of the entropy of the gas can easily be understood by referring to the connection between entropy and probability suggested by Boltzmann. Because it is less probable that the gas particles will occupy a smaller volume, such a state has a higher order, and hence lower entropy. Interestingly, for the case of radiation filling a cavity, Einstein merely uses the Wien black-body radiation density, which is known to be correct only for high radiation frequencies.

Einstein’s crucial insight comes when he observes that the entropy of light in a cavity varies in exactly the same way with the volume of the cavity as the entropy of a gas. On the basis of this observation, he suggests that light also consists of particles which he calls light quanta. He clearly states that this is only a heuristic point of view and not a logically binding conclusion. Only in the last chapter (of eight) of the paper does Einstein finally get to the photoelectric effect by asking where quanta of light might have implications. He notes that it would naturally explain why the wavelength of light emitted in photoluminescence is always larger than that of the absorbed light. This is because a single particle of light is absorbed and unless additional energy is supplied, the energy of the emitted particles of light in general is lower.

When coming finally to the photoelectric effect Einstein observes that the energy of the emitted electrons, as measured by Lenard⁸⁴, can be understood quantitatively by means of his light quanta (see also Box 2). The only, but crucial, prediction he makes is that the maximum energy of the electrons must vary linearly with the frequency of the incident light. This prediction was confirmed experimentally to high precision ten years later by Millikan⁸⁵. Millikan could extract from the slope of his measured curve a value for Planck’s constant h that precisely agreed with the number found in earlier measurements of the black-body radiation. This is one of the most convincing confirmations of the idea of quanta. Millikan recalled: “I spent ten years of my life

testing that 1905 equation of Einstein’s and, contrary to my expectations, I was compelled in 1915 to assert its unambiguous experimental verification in spite of its unreasonableness since it seemed to violate everything that we knew about the interference of light”⁸⁶. Indeed, while Millikan proved the validity of Einstein’s equation beyond doubt, he categorically rejected Einstein’s light-quantum hypothesis as an interpretation of it. Only after the discovery of the Compton effect in 1923 (ref. 87) and subsequent experiments⁸⁸ did Millikan, like many other physicists, accept Einstein’s light-quantum hypothesis. These experiments established the conservation of energy and momentum of individual light quanta for the specific case of elastic scattering from electrons, and so finally made it clear that Einstein’s 1905 conception was more than simply “heuristic”⁸⁹.

However, the apparent conflict between a corpuscular theory and interference could not be resolved before quantum mechanics itself was fully developed. Once Schrödinger’s and Heisenberg’s formulations of quantum mechanics were known, it was obvious that these should be applied to the electromagnetic cavity oscillators and eventually to the field itself⁷⁹. QED accommodates both interference and quantization. Its fields are built on Maxwell’s equations and populated with integral numbers of photons. QED gives us the fundamental properties of the photon: the photon has no rest mass, or, equivalently, moves at the vacuum speed of light. Any finite rest mass would make the vacuum dispersive and modify Coulomb’s law. Experiments put an upper limit of about 10^{-50} kg on the photon mass⁹⁰. The photon is also predicted to have zero electric charge. The experimental upper limit is approximately 10^{-17} e (ref. 91).

Einstein’s observations about the entropy of radiation are intimately connected to the quantum statistics of photons, and from Bose’s and Einstein’s work we know that the photon is a boson. In accordance with the vector character of the electromagnetic field it must therefore have spin 1. Further, equivalent to the transversality of electromagnetic radiation and as a consequence of the photon’s zero rest mass, we know that only the spin eigenstates of $+1$ and -1 along its linear momentum are allowed. These give rise to the two orthogonal polarizations of light. Finally, in modern field theory the photon is the exchange particle of electromagnetic interaction and it was the first example of a gauge boson, eventually leading to the gauge theories that form today’s standard model⁹².

Box 2

Semiclassical radiation theories are a dead end

A fascinating irony is that the photoelectric effect, as it was known in Einstein’s time, can be understood without having to assume that light, or, in modern terms, the radiation field, is quantized. It suffices to assume that the surface can only absorb or emit light in energy quanta. More generally speaking, many phenomena thought to be due to the quantum nature of light can actually be explained by using a classical electromagnetic field and by assuming that only the processes of absorption and emission are quantized. In the simplest way, this is done by assuming that the absorbers consist of oscillators which can absorb and emit radiation in quantized packets only. Among the initial advocates of this semiclassical theory, in which only the atoms are quantized while the electromagnetic field remains as classical waves, were Planck and Bohr themselves.

Bohr even pursued a wave-theoretic explanation of the Compton effect within the later refuted Bohr–Kramers–Slater (BKS)

theory. It turns out, however, that semiclassical ideas cannot account for all experimental observations. Examples are experiments showing that there is no lower limit on the accumulation time of radiation energy in the photoelectric effect, which suggests an instantaneous energy transfer, such as would be expected from a particle-like interaction^{93,94}. Other examples are correlation experiments related to quantum entanglement, which can in principle not be modelled by local classical theories.

Curiously, Einstein together with Podolsky and Rosen first discussed such correlations in 1935 (ref. 22) for completely different reasons and apparently without perceiving that this famous paper would eventually deliver another independent proof of the photon hypothesis. A detailed discussion of the quantum versus the semiclassical approach in the light of quantum non-locality has been given by Clauser⁹⁵. The interested reader is referred to the accounts of Klein⁹⁶, Stachel⁹⁷, Pais⁹⁸ or Clauser⁹⁹.

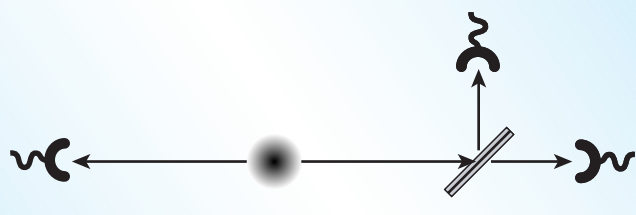


Figure 1 Principle of Clauser's experiment with correlated pairs of photons (simplified). The source emits two photons. Registration of a photon on the left detector provides the information that one and only one photon at the right side encounters a 50/50 beam splitter where it is either reflected or transmitted. The fact that only one of the two detectors behind the beam splitter registers and never both can easily be understood using the photon concept, and is in clear conflict with semiclassical theories of radiation⁴.

finding that each individual photon 'knows' it should never end up in the minimum of an interference fringe.

We emphasize that the conceptual questions arising for photon interference are the same as those arising for interference of massive particles. In both cases we see particle-like and wave-like properties. An inequivalence arises for certain interference experiments¹⁴ because the photon has no rest mass.

Two-photon interference

An interesting consequence of the bosonic character of photons is their bunching behaviour. This is seen most directly when two photons — one from each input port — are incident on a beam splitter (Fig. 3). If the two photons do not arrive simultaneously, each has a 50% chance of going either way after the beam splitter, independently of the other photon. This results in the coincidences shown. But if the photons arrive simultaneously, they become indistinguishable and end up together randomly in either beam. In the experiment the rate of coincident photon detections at the beam splitter outputs is monitored. The resulting dip in the coincidence rate is called the Hong–Ou–Mandel dip¹⁵.

What happens is a quantum interference effect. The only way that one photon can arrive at each detector is if both photons are either reflected or transmitted. The detection probability in quantum physics is given by the square of the probability amplitude, which is different from squaring the actual electromagnetic field. Curiously, the probability amplitudes for these two possibilities destructively interfere with each other. This results from the well-known phase jump of 90 degrees that each photon experiences upon reflection. This implies a total phase of 180 degrees of the state $|\text{both photons reflected}\rangle$ relative to the state $|\text{both photons transmitted}\rangle$.

Fermions would behave differently because their quantum state is antisymmetric, as reflected by a negative sign in their initial state. In this case the two amplitudes introduced above interfere constructively and the two particles are always found in separate outputs. Interestingly, this 'fermionic' behaviour can also be observed for two photons if the photons are prepared in an antisymmetric state with respect to their spin (Fig. 3). This latter observation turned out to be crucial for many quantum information applications, specifically for quantum dense coding¹⁶.

Complementarity, information and quantum physics

Complementarity, the mutually exclusive nature of the wave and particle concepts, has led to intense discussions. Of these, the early ones between Einstein and Bohr raised the key issues. Whereas Einstein thought that it should be possible to observe an interference pattern and at the same time know for each photon which slit it went through, Bohr was always able to show that an apparatus capable of determining the particle's path was by necessity constructed such

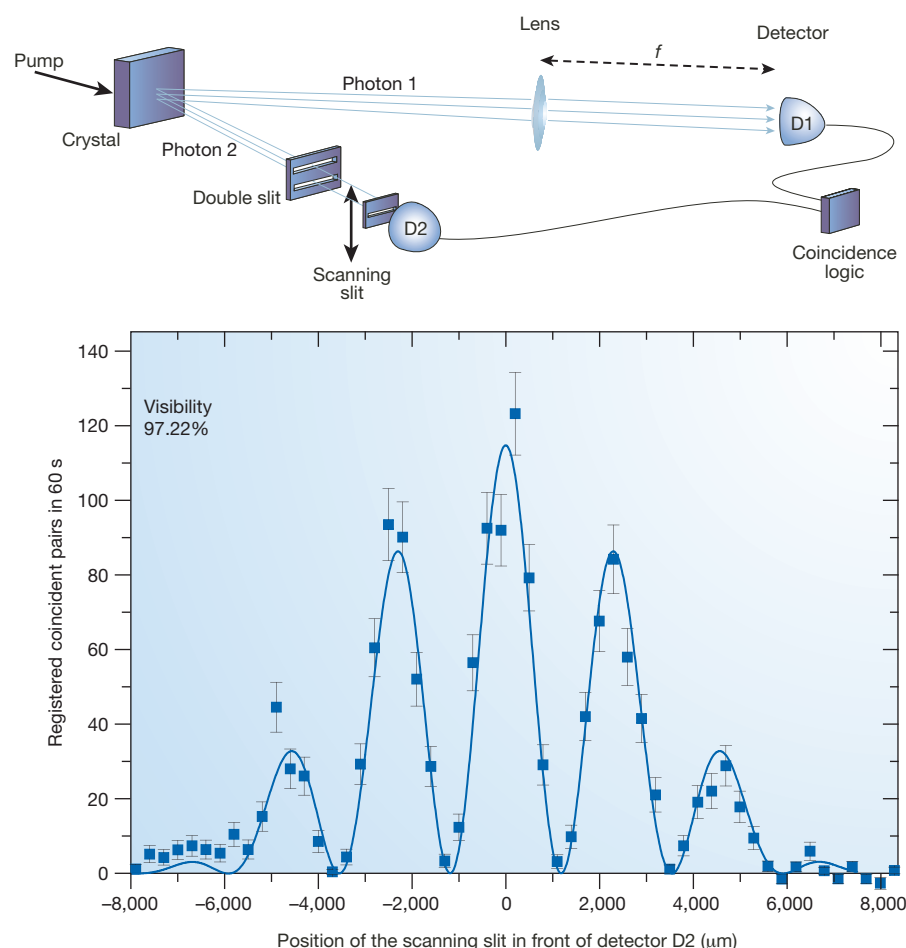


Figure 2 Single-photon double-slit interference. A pair of momentum-entangled photons is created by type-I parametric down-conversion. Photon 2 enters a double-slit assembly and photon 1 is registered by a detector D1 placed at distance f in the focal plane of the lens. This projects the state of photon 2 into a momentum eigenstate which cannot reveal any positional information and, hence, supplies no information about slit passage. Therefore, in coincidence with a registration of photon 1 in the focal plane, photon 2 exhibits the interference pattern shown. On the other hand, when the detector is placed in the imaging plane, it does reveal the path photon 2 takes through the slit assembly, which therefore does not show the interference pattern. The observed count rate of at most two photons per second implies that the average spatial distance between photons registered would be of the order of 100,000 km or more. Therefore, most of the time the apparatus is empty (from refs 11 and 12). The error bars (s.d.) show the statistical errors of photon counting.

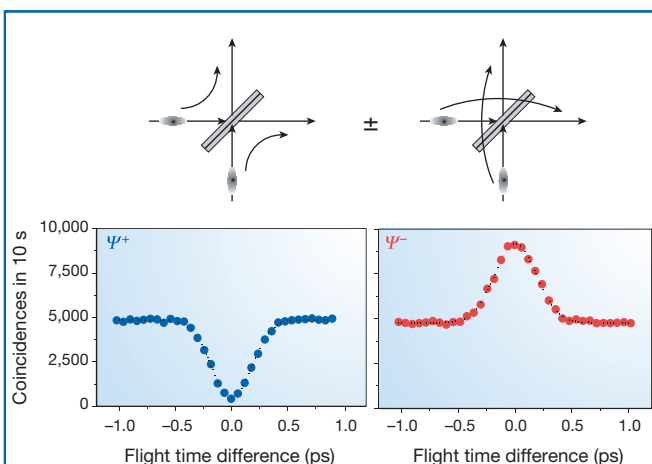


Figure 3 Bunching (left) or antibunching (right) behaviour of photon pairs. One photon each is incident from each input of a 50/50 beam splitter. Coincidences between detectors in the two output beams are registered as a function of the flight time difference of the incident photons. No coincidences are observed for zero flight time difference (left) for the usual symmetric spatial state of the two photons. This is because the probability amplitudes for the transmission of both photons and for the reflection of both photons destructively interfere: the latter one picks up a minus sign owing to the phase shift of the photons upon reflection. Interestingly, the two incident photons can also be in an antisymmetric spatial state (which occurs if the two-photon spin state is also antisymmetric). In this case, the two amplitudes interfere constructively. This results in the two photons always exiting in separate beams for zero flight time difference. The observed coincidence peak (right) confirms this expected antibunching.

that no interference pattern could arise and vice versa¹⁷. Today, it is thought that a perfect interference pattern arises only when there is no possible way of finding out which path the particle took. Evidently, intermediate cases are also possible, of partial path information together with non-perfect interference fringes¹⁸. In so-called delayed choice experiments^{19,20} the decision of whether to observe path information can be delayed to when the particle is already inside the interferometer setup and even until after it has been registered. This again supports the view that the quantum state may be seen as a representation of the information about the probabilities of possible measurement results, which may include mutually exclusive, that is, complementary ones.

An experiment supporting the information aspect of quantum interference²¹ has been performed by Mandel's group at Rochester as part of a series of ground-breaking experiments on the quantum nature of light. In their experiment (Fig. 4), they used the emission of one photon pair from two down-conversion crystals. One photon passed a modified Mach–Zehnder interferometer, and what happened to the other photon decided whether the first photon showed interference or not. Thus, the still widespread view that the act of determining the path taken by the particle disturbs its state enough to destroy the interference is untenable. The key factor is whether path information is available: it does not matter if someone takes care to read it out or not. Indeed, there have even been experiments where the path information carried by the second particle is destroyed after the particle passing through the double slit has already been registered. Here, the interference pattern is still observed. The double-slit diffraction pattern shown in Fig. 2 was obtained in this way.

When analysing quantum interference we can fall into all kinds of traps. The general conceptual problem is that we tend to reify — to take too realistically — concepts like wave and particle. Indeed if we consider the quantum state representing the wave simply as a calculational tool, problems do not arise. In this case, we should not talk about a wave propagating through the double-slit setup or through a Mach–Zehnder interferometer; the quantum state is simply a tool to calculate probabilities. Probabilities of the photon being somewhere? No, we should be even more cautious and only talk about probabilities of a photon detector firing if it is placed somewhere. One might be tempted, as was Einstein², to consider the photon as being localized at some place with us just not knowing that place. But, whenever we talk about a particle, or more specifically a photon, we should only mean that which a 'click in the detector' refers to.

Nonlocality, Bell and GHZ

Einstein did not only criticize quantum mechanics for the new role of randomness mentioned above. His criticism went much further in his insistence on the existence of a real factual world and on the role of physics to describe that reality. This criticism forms the basis of his famous article published in 1935, together with Podolsky and Rosen. The Einstein–Podolsky–Rosen (EPR) paper²² makes use of correlations shown by entangled quantum states. Bell discovered in 1964 (ref. 23) that the predictions of quantum physics for these correlations are at variance with a local realistic world view. Such a classically intuitive view holds that the outcome of a measurement on a physical system is determined by physical properties of the system prior to and independent of the measurement (realism), and that the outcome cannot depend on any actions in space-like separated regions (Einstein locality).

The quantum correlations are too strong to be reproduced by any such model. After the initial experiments by Freedman and Clauser²⁴, and the much refined experiments by Aspect²⁵ (Fig. 5) confirming the quantum predictions, two loopholes remained open. Thus, a local realistic viewpoint remained at least logically possible. The first loophole, the so-called communication loophole, used the fact that even before the two photons are registered, the apparatus settings could be communicated to detectors and/or to the source. A beautiful experiment to close this loophole using periodically switched time-dependent polarizer settings was performed by Aspect in 1982 (ref. 26). The loophole was more decisively ruled out by an experiment of Weihs *et al.*²⁷ where the measurement settings

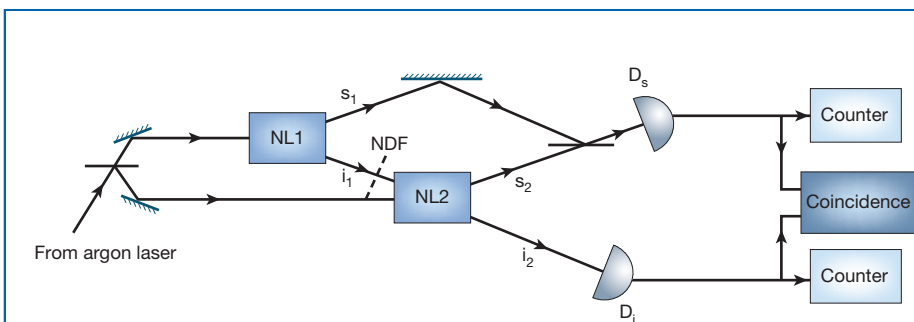


Figure 4 A mind-boggling interference experiment (from ref. 21). Two nonlinear crystals NL1 and NL2 are pumped by the same laser and produce one pair of photons in the superposition of being in beams s_1 and i_1 , or in beams s_2 and i_2 . After transmission through NL2, the photons from i_1 are indistinguishable from photons in i_2 and thus the original distinguishability of the path of s_1 and s_2 disappears, and quantum interference is observed for pairs of photons that are detected in coincidence between detectors D_i and D_s . Insertion of an absorber (neutral density filter, NDF) can be used gradually to introduce distinguishability and thus to make the interference disappear. Note that no disturbance whatsoever acts on the interfering photon in beams s_1 and s_2 .

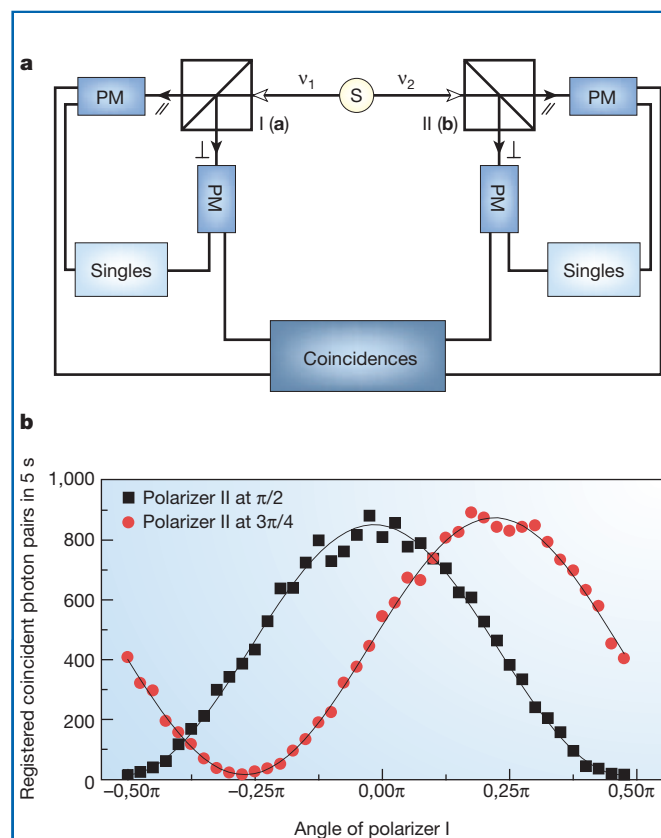


Figure 5 Bell tests violating local realistic predictions. **a**, Setup of the experimental test of Bell's inequality with correlated photon pairs produced by atomic cascade relaxation (from ref. 25). The two photons v_1 and v_2 are analysed in separate polarizers (I in direction **a** and II in direction **b**), and detected with photomultipliers (PM). A dedicated electronic circuit is used to determine the coincidences, which are included in evaluating the correlation terms in Bell's inequality. This and several other similar experiments proved with high significance that real particles follow the prediction of quantum mechanics, and violate the limits imposed by local realistic theories. **b**, Correlation curves (from ref. 27) in a Bell inequality experiment. A characteristic of such experiments is that the correlation depends on the difference angle of the analysers only and not on any of their absolute values.

were changed randomly and fast with respect to the distance between the stations. This left only the so-called detection loophole open, which implied that although all photon pairs would obey local realism and hence be at variance with quantum physics, the small detected subset confirmed quantum mechanics. Since the detection efficiencies were far from ideal, such reasoning could not be ruled out. This latter loophole was closed in an experiment on entanglement between two ions in a cavity²⁸, in which it was possible to detect nearly all entangled pairs. So, although both remaining loopholes have now been closed in separate experiments, the logical possibility exists that nature tricks us and makes use of different loopholes in different experiments. Although no one reasonably assumes nature to be so capricious, a future experiment that closes both loopholes together would still be interesting.

The conflict of local realism with quantum mechanics, first exposed by Bell for entangled pairs, is even more striking for three or more entangled particles. For the so-called GHZ (Greenberger–Horne–Zeilinger) states^{29,30}, situations exist for which a local realist and a quantum mechanic make completely opposite predictions, even for individual measurement results on one photon. Needless to say, experiments³¹ have confirmed the quantum prediction (Fig. 6).

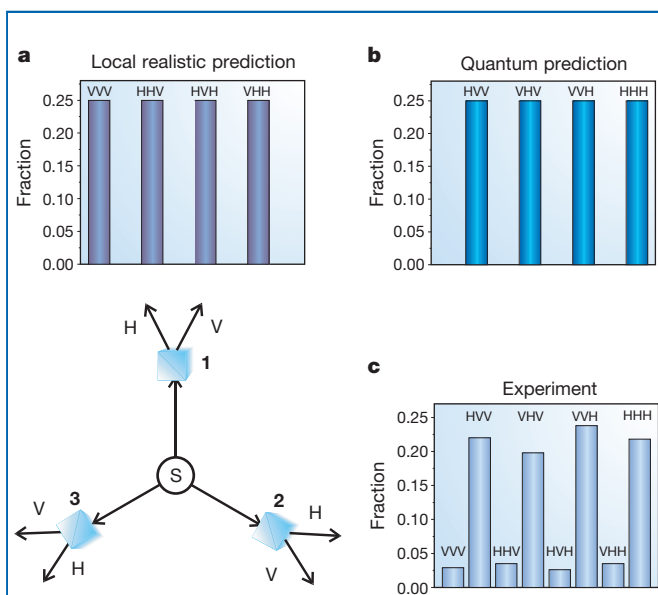


Figure 6 Sketch of an experimental test of the GHZ argument against local realistic theories (from ref. 31). A source (S) emits polarization entangled photon triplets in a GHZ state (bottom left). Measurements in three different combinations of circular and diagonal polarization (not shown) show perfect correlation. These correlations define local realistic elements of reality, which lead to definite predictions (**a**) for a measurement in all diagonal polarization bases. The quantum predictions (**b**) are exactly opposite. **c**, The experimental relative frequencies of triple coincidences behind polarizers 1, 2 and 3. For example, for photons 2 and 3, which are both horizontally (H) polarized, local realism (**a**) predicts the first photon to be vertically (V) polarized, quantum mechanics (**b**) predicts it to be H polarized, and experiment (**c**) confirms the predictions of quantum mechanics within experimental accuracy.

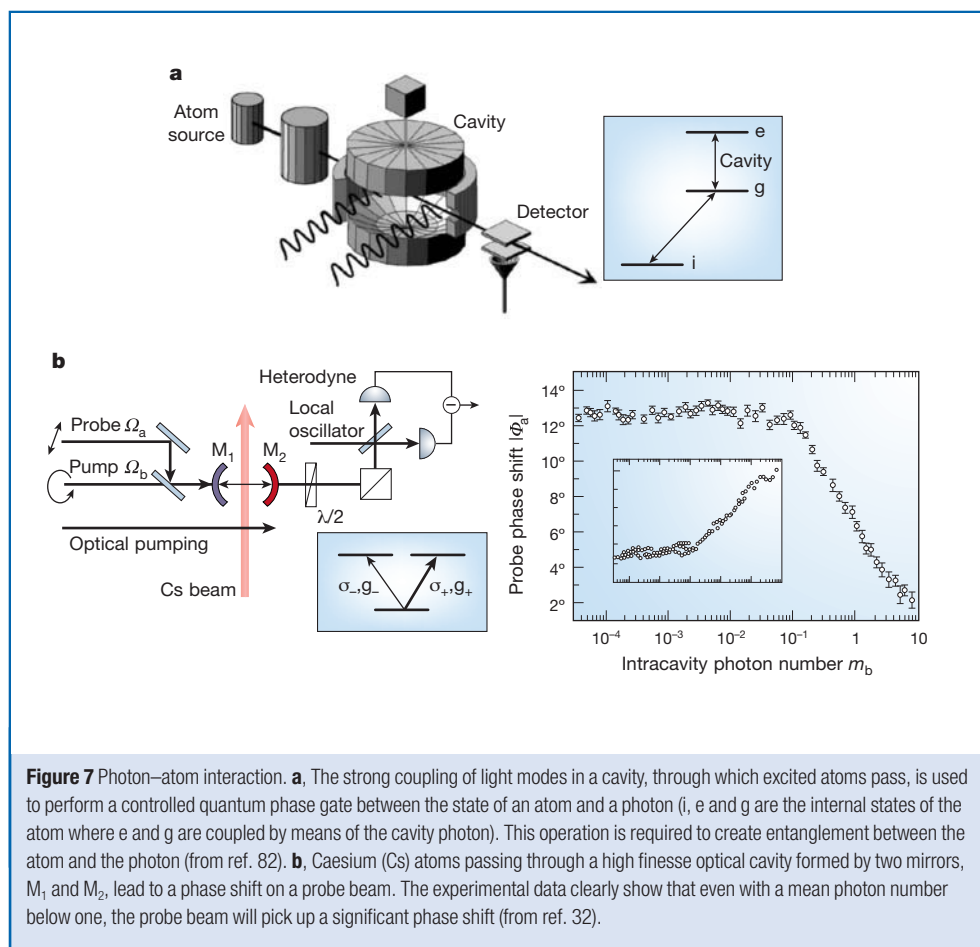
Photons, atoms and beyond

As mentioned above, the initial question of whether it is only matter or also radiation that is quantized has finally been settled in favour of the photon. It has now become possible to investigate the interaction between photons and atoms in great detail. For example, Kimble's group at Caltech³² showed that it was possible to observe the phase shift experienced by an atom while it interacted with a field of, on average, less than one photon. Haroche and his group at the École Normale in Paris³³ were able to construct entangled states between single photons trapped in a high-finesse cavity and atoms passing through (Fig. 7). Such experiments have also been used to demonstrate several interesting aspects, such as time-resolved quantum interference phenomena³⁴, trapping of atoms with single photons^{35,36} or quantum non-demolition measurements, in which the presence of single photons can be determined without destroying the photon³⁷.

Quantum information processing with photons

In the emerging field of quantum information technology the two basic subfields are quantum communication and quantum computation. The photon has been put to work in recent years, particularly in new concepts of communication. In quantum cryptography the complementarity of different measurements on a quantum system^{38–40} is used to establish a secure key between two partners. In quantum teleportation^{41,42} it is possible to transfer the quantum state of independent particles from one system to another by employing entangled states as a quantum information channel.

When turning to future challenges and developments, it is likely that the photon will have a significant role in quantum communication. The currently most advanced source for entanglement, an important resource for quantum information processing, is photon



that some gates could be realized through teleportation⁵⁸, an important breakthrough was the suggestion by Knill *et al.*⁵⁹ that even with linear optical elements, universal quantum computation could be realized. Following these suggestions, various quantum computation primitives have been demonstrated with photons alone — including conditional phase shift operations⁶⁰, and destructive^{61,62} and even non-destructive controlled NOT (CNOT) gates⁶³. All these schemes use unentangled states as inputs on which the quantum gates operate. A new and probably more practical approach is the concept of a ‘one-way quantum computer’⁶⁴ that realizes universal quantum computation in a way that is totally different from that used by existing quantum computing schemes. Here, the idea is to start with a general, highly entangled multi-qubit state. The computation is then performed by applying a sequence of simple one-particle measurements, specific to the algorithm implemented. This new approach uses highly entangled cluster states, which recently have been realized with photons and applied to demonstrate elementary quantum gates⁶⁵. Entangled multi-particle states also have a significant

role in other new protocols of quantum information. For example, quantum error correction is based on such states^{66,67}.

Many laboratories all over the world are working towards developing many different physical implementations both of quantum communication devices and of quantum computers, and it will be interesting to see which technology will be the best. Yet, we are convinced that some day in the future, the present classical information technology will be replaced by a quantum one, even if this is only because of the continuing miniaturization of switching elements in computer chips.

For future technological developments, new sources for single-photon states will be needed. The most basic of such sources would be a single-photon source that, on demand, produces one, and only one, photon at a specific time and not at random. There has been important progress over the past few years in this field from various directions, including atoms in cavities^{68,69} and solid-state devices such as cavity-coupled quantum dots⁷⁰. An extensive account of such activities has recently been collected by Grangier *et al.*⁷¹. More generally, it would be good to have sources that produce any specific multiphoton state, even entangled ones, on demand. Promising experiments along this line have been performed in the context of cavity quantum electrodynamics (QED; ref. 72). Also, more efficient photon detectors are needed that operate over a broader wavelength range than those currently available. A particularly interesting development would be a detector that is able to discriminate clearly between one and two photons have been reported, for example, by Yamamoto’s group⁷³, by using a visible-light photon counter (VLPC). In the far future, a detector that identifies deterministically an arbitrary N -photon state would be useful.

We have been able to give only a glimpse of the vast expanse of implications and applications of the photon concept, and of quantum

pairs generated in SPDC (ref. 43). Moreover, only with photons is it possible to cover large distances outside the protected environment of the laboratory. There have been experiments transporting entangled states over more than 10 km using glass fibres⁴⁴ and across the river Danube in free space⁴⁵. Quantum cryptography, with faint laser pulses containing less than one photon on average, has been tested in free space for distances of more than 20 km (ref. 46) — even in daylight⁴⁷ — and in optical fibres for a physical separation of 67 km (ref. 48). Furthermore, quantum communication through satellites is the only possible way to cover global distances. Satellite-based quantum communication may very well be realized within the next decade.

In terms of technical applications of the photon idea, the most advanced is quantum cryptography^{49,50} (Fig. 8). Prototype devices are already on the market and the development of systems that are suitable for the security industry is well under way. Quantum teleportation might one day provide useful communication links between yet-to-be-developed quantum computers. Initial tests of long-distance quantum teleportation have recently been performed in Geneva⁵¹ and in Vienna⁵². The latter was a real field test that even included active feed-forward of measurement results (Fig. 9). An important extension of teleportation in this sense will be quantum repeaters⁵³, a combination of entanglement swapping (that is, the teleportation of an entangled state)^{54,55} and local atomic memories of quantum information, which also exploit the atom-photon interface^{56,57}. The development of these applications is intimately connected to the development of the quantum computer itself.

Although for quantum communication the obvious choice is photons, for quantum computation, implementations in localized systems like atoms, ions or solid-state devices seem to be preferable. Yet, surprisingly, even for implementing quantum computation algorithms, photons offer interesting possibilities, despite the considerable difficulty of storing them for a long time. After the discovery

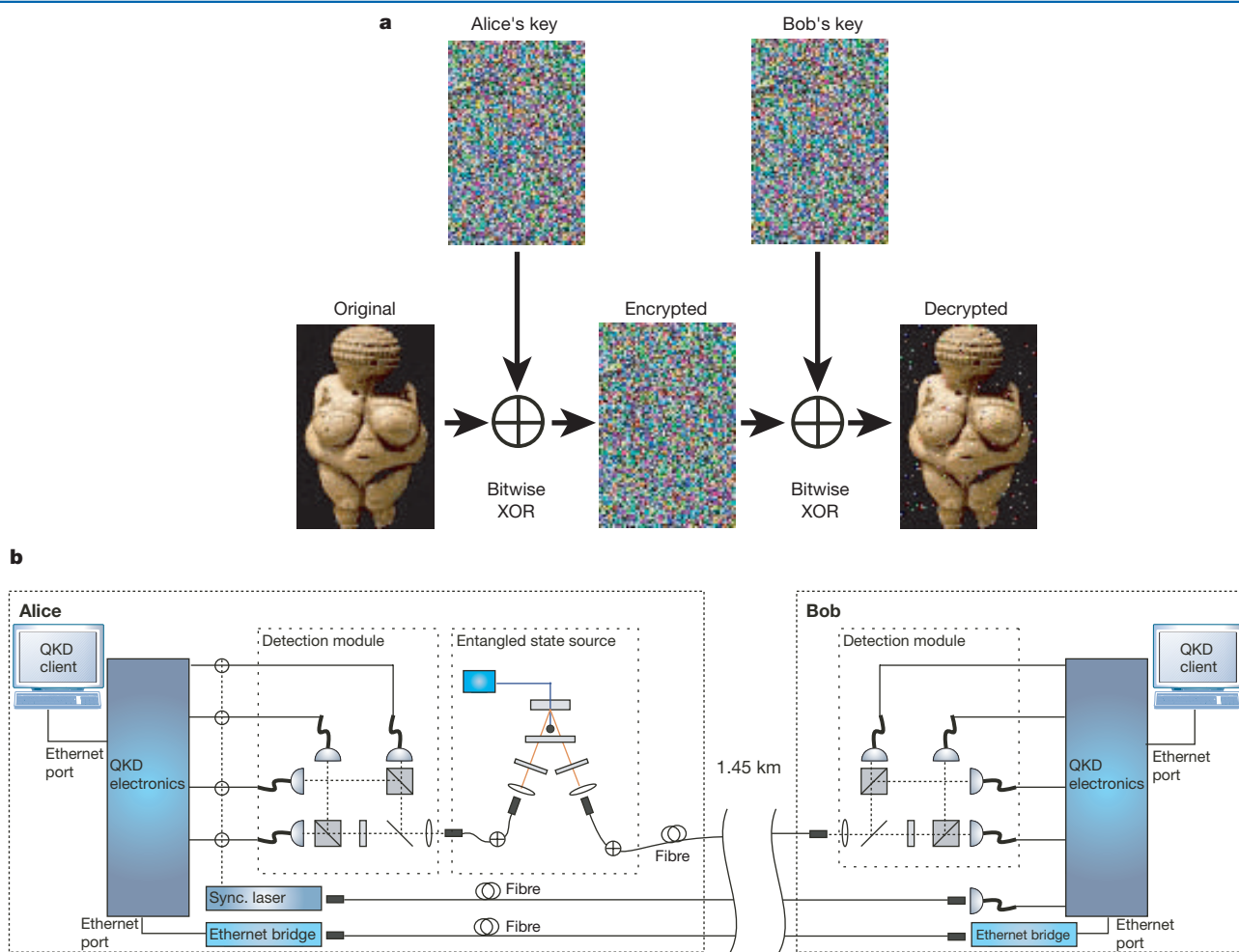


Figure 8 Quantum cryptography in practice. Entangled pairs of photons are used to generate secure keys at Alice's and Bob's distant stations. The local measurement outcomes at Alice are completely random but correlated to the results at Bob. Thus, a cryptographic key is created whose security rests only on the principles of quantum

physics. **a**, The secure transmission of an image of the 'Venus of Willendorf' over 350 m (ref. 83). **b**, Experimental setup of a prototype system for entangled photon quantum key distribution (QKD), which has even been used in a real world demonstration — to securely transfer money into a bank account (from ref. 50).

optics in general. Also deserving a mention is the wide field of experiments on squeezed states initialized by Slusher *et al.*⁷⁴ and rapidly expanded by others. These include continuous-variable demonstrations of quantum teleportation⁷⁵, of quantum optics in phase

space^{76,77}, and of quantum cryptography⁷⁸. Here, phenomena are studied that are a consequence of the quantization of the electromagnetic field⁷⁹. But in general, the concept of the photon as an individual particle is less important here.

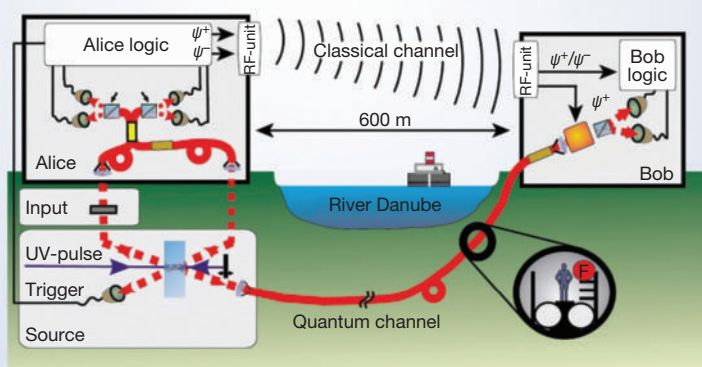


Figure 9 Schematic setup of quantum state teleportation across the river Danube (from ref. 52). The polarization state of a photon is transferred from Alice to Bob through the use of a quantum channel (entangled photons) and a classical channel (microwave pulses). Bob was able to regain the original photon state with a fidelity as high as 90%, which is clearly above the limits imposed by classical concepts.

Conclusion

Evidently, Einstein's 1905 proposal of the photon concept has had tremendous impact. But Einstein should also be highly credited for his various criticisms of quantum physics that were part of the early debate with his contemporaries (including Bohr). They triggered a body of both theory and experiments concerned with individual quantum systems. In this context, experiments with photons have had a pioneering role. Although such experiments now rule out Einstein's point of view, they gave rise to the new fields of quantum information processing. But the conceptual problems are not fully settled. This is signified by the wide spectrum of different interpretations of quantum physics that compete with each other. In our view, a common trait of many interpretations is that entities are taken to be 'real' beyond necessity. This is most obvious for the case of the 'many-worlds interpretation'⁸⁰ where the coexistence of parallel worlds is

claimed without compelling evidence, but it also holds, for example, for the Bohm interpretation⁸¹ where, again without compelling evidence, each particle is given a well-defined position and momentum at any time. We suggest that these are simply attempts to keep, in one way or other, a realistic view of the world. It may well be that in the future, quantum physics will be superseded by a new theory, but it is likely that this will be much more radical than anything we have today. □

doi:10.1038/nature03280

1. Einstein, A. Zur Elektrodynamik bewegter Körper. *Ann. Phys.* **17**, 891–921 (1905).
2. Einstein, A. Über einen die Erzeugung und Verwandlung des Lichtes betreffenden heuristischen Gesichtspunkt. *Ann. Phys.* **17**, 132–148 (1905).
3. Lewis, G. N. The conservation of photons. *Nature* **118**, 874–875 (1926).
4. Clauser, J. F. Experimental distinction between the classical and quantum field-theoretic predictions for the photoelectric effect. *Phys. Rev. D* **9**, 853–860 (1974).
5. Burnham, D. C. & Weinberg, D. L. Observation of simultaneity in parametric production of optical photon pairs. *Phys. Rev. Lett.* **25**, 84–87 (1970).
6. Jennewein, T., Achleitner, A., Weihs, G., Weinfurter, H. & Zeilinger, A. A fast and compact quantum random number generator. *Rev. Sci. Instrum.* **71**, 1675–1680 (2000).
7. Kimble, H. J., Dagenais, M. & Mandel, L. Photon antibunching in resonance fluorescence. *Phys. Rev. Lett.* **39**, 691–695 (1976).
8. Diedrich, F. & Walther, H. Nonclassical radiation of a single stored ion. *Phys. Rev. Lett.* **58**, 203–206 (1987).
9. Taylor, G. I. Interference fringes with feeble light. *Proc. Camb. Phil. Soc. Math. Phys. Sci.* **15**, 114–115 (1909).
10. Grangier, P., Roger, G. & Aspect, A. Experimental evidence for a photon anticorrelation effect on a beam splitter: a new light on single-photon interference. *Europhys. Lett.* **1**, 173–179 (1986).
11. Dopfer, B. *Zwei Experimente zur Interferenz von Zwei-Photon Zuständen: Ein Heisenbergmikroskop und Pendellösung*. PhD thesis, Universität Innsbruck (1998).
12. Zeilinger, A. Experiment and the foundations of quantum physics. *Rev. Mod. Phys.* **71**, S288–S297 (2000).
13. Feynman, R. P., Leighton, R. B. & Sands, M. L. *The Feynman Lectures on Physics* Vol. 3 (Addison-Wesley, Massachusetts, 1965).
14. Brukner, C. & Zeilinger, A. Nonequivalence between stationary matter wave optics and stationary light optics. *Phys. Rev. Lett.* **79**, 2599–2603 (1997).
15. Hong, C. K., Ou, Z. Y. & Mandel, L. Measurement of subpicosecond time intervals between two photons by interference. *Phys. Rev. Lett.* **59**, 2044–2046 (1987).
16. Mattle, K., Weinfurter, H., Kwiat, P. G. & Zeilinger, A. Dense coding in experimental quantum communication. *Phys. Rev. Lett.* **76**, 4656–4659 (1996).
17. Bohr, N. in *Albert Einstein: Philosopher-Scientist VII* (ed. Schilpp, P. A.) Vol. 7, 201–241 (Library of Living Philosophers, Evanston, 1949).
18. Englert, B.-G. Fringe visibility and which-way information: An inequality. *Phys. Rev. Lett.* **77**, 2154–2157 (1997).
19. Wheeler, J. A. in *Problems in the Foundations of Physics*. *Proc. Intl. School of Physics 'Enrico Fermi', Course LXXII*. (ed. di Francia, G. T.) 395–492 (North-Hollands, Amsterdam, 1979).
20. Scully, M. O. & Drühl, K. Quantum eraser: A proposed photon correlation experiment concerning observation and “delayed choice” in quantum mechanics. *Phys. Rev. A* **25**, 2208–2213 (1982).
21. Zou, X. Y., Wang, L. J. & Mandel, L. Induced coherence and indistinguishability in optical interference. *Phys. Rev. Lett.* **67**, 318–321 (1991).
22. Einstein, A., Podolsky, B. & Rosen, N. Can quantum mechanical description of physical reality be considered complete? *Phys. Rev.* **47**, 777–780 (1935).
23. Bell, J. S. On the Einstein Podolsky Rosen paradox. *Physics* **1**, 195–200 (1964).
24. Freedman, S. J. & Clauser, J. F. Experimental test of local hidden-variable theories. *Phys. Rev. Lett.* **28**, 938–941 (1972).
25. Aspect, A., Grangier, P. & Roger, G. Experimental realization of Einstein-Podolsky-Rosen-Bohm Gedankenexperiment: A new violation of Bell's inequalities. *Phys. Rev. Lett.* **49**, 91–94 (1982).
26. Aspect, A., Dalibard, J. & Roger, G. Experimental test of Bell's inequalities using time-varying analyzers. *Phys. Rev. Lett.* **49**, 1804–1807 (1982).
27. Weihs, G., Jennewein, T., Simon, C., Weinfurter, H. & Zeilinger, A. Violation of Bell's inequality under strict Einstein locality conditions. *Phys. Rev. Lett.* **81**, 5039–5043 (1998).
28. Rowe, M. A. et al. Experimental violation of a Bell's inequality with efficient detection. *Nature* **409**, 791–794 (2001).
29. Greenberger, D. M., Horne, M. A. & Zeilinger, A. in *Bell's Theorem, Quantum Theory, and Conceptions of the Universe* (ed. Kafatos, M.) 73–76 (Kluwer, Dordrecht, 1989).
30. Greenberger, D. M., Horne, M. A., Shimony, A. & Zeilinger, A. Bell's theorem without inequalities. *Am. J. Phys.* **58**, 1131–1143 (1990).
31. Pan, J.-W., Bouwmeester, D., Daniell, M., Weinfurter, H. & Zeilinger, A. Experimental test of quantum nonlocality in three-photon Greenberger-Horne-Zeilinger entanglement. *Nature* **403**, 515–518 (2000).
32. Turchette, Q. A., Hood, C. J., Lange, W., Mabuchi, H. & Kimble, H. J. Measurement of conditional phase shifts for quantum logic. *Phys. Rev. Lett.* **75**, 4710–4713 (1995).
33. Rauschenbeutel, A. et al. Step-by-step engineered multiparticle entanglement. *Science* **288**, 2024–2028 (2000).
34. Legero, T., Wilk, T., Kuhn, A. & Rempe, G. Time-resolved two-photon quantum interference. *Appl. Phys. B* **77**, 797–802 (2003).
35. Pinkse, P., Fischer, T., Maunz, P. & Rempe, G. Trapping an atom with single photons. *Nature* **404**, 365–368 (2000).
36. Hood, C. J., Lynn, T. W., Doherty, A. C., Parkins, A. S. & Kimble, H. J. The atom-cavity microscope: Single atoms bound in orbit by single photons. *Science* **287**, 1447–1453 (2000).
37. Nogues, G. et al. Seeing a single photon without destroying it. *Nature* **400**, 239–242 (1999).
38. Wiesner, S. Conjugate coding. *SIGACT News* **15**, 78–88 (1983).
39. Bennett, C. H. & Brassard, G. in *Proc. IEEE Intl. Conf. on Computers, Systems and Signal Processing* 175–179 (IEEE, Bangalore, 1984).
40. Ekert, A. K. Quantum cryptography based on Bell's theorem. *Phys. Rev. Lett.* **67**, 661–663 (1991).
41. Bennett, C. H. et al. Teleporting an unknown quantum state via dual classical and Einstein-Podolsky-Rosen channels. *Phys. Rev. Lett.* **70**, 1895–1899 (1993).
42. Bouwmeester, D. et al. Experimental quantum teleportation. *Nature* **390**, 575–579 (1997).
43. Kwiat, P. G. et al. New high-intensity source of polarization-entangled photon pairs. *Phys. Rev. Lett.* **75**, 4337–4341 (1995).
44. Tittel, W., Brendel, J., Zbinden, H. & Gisin, N. Violation of Bell inequalities by photons more than 10 km apart. *Phys. Rev. Lett.* **81**, 3563–3566 (1998).
45. Aspelmeyer, M. et al. Long-distance free-space distribution of quantum entanglement. *Science* **301**, 621–623 (2003).
46. Kurtsiefer, C. et al. A step towards global key distribution. *Nature* **419**, 450 (2002).
47. Hughes, R. J., Nordholt, J. E., Derkacs, D. & Peterson, C. G. Practical free-space quantum key distribution over 10 km in daylight and at night. *New J. Phys.* **4**, 43, 1–43 (2002).
48. Stucki, D., Gisin, N., Guinnard, O., Ribordy, G. & Zbinden, H. Quantum key distribution over 67 km with a plug&play system. *New J. Phys.* **4**, 1–41 (2002).
49. Gisin, N., Ribordy, G., Tittel, W. & Zbinden, H. Quantum cryptography. *Rev. Mod. Phys.* **74**, 145–195 (2002).
50. Poppe, A. et al. Practical quantum key distribution with polarization entangled photons. *Opt. Express* **12**, 3865–3871 (2004).
51. Marcikic, L., de Riedmatten, H., Tittel, W., Zbinden, H. & Gisin, N. Long-distance teleportation of qubits at telecommunication wavelengths. *Nature* **421**, 509–513 (2003).
52. Ursin, R. et al. Quantum teleportation across the Danube. *Nature* **430**, 849 (2004).
53. Briegel, H.-J., Dür, W., Cirac, J. I. & Zoller, P. Quantum repeaters: The role of imperfect local operations in quantum communication. *Phys. Rev. Lett.* **81**, 5932–5935 (1998).
54. Zukowski, M., Zeilinger, A., Horne, M. A. & Ekert, A. K. “Event-ready-detectors” Bell experiment via entanglement swapping. *Phys. Rev. Lett.* **71**, 4287–4290 (1993).
55. Jennewein, T., Weihs, G., Pan, J.-W. & Zeilinger, A. Experimental nonlocality proof of quantum teleportation and entanglement swapping. *Phys. Rev. Lett.* **88**, 017903 (2002).
56. Blinov, B. B., Moehring, D. L., Duan, L.-M. & Monroe, C. Observation of entanglement between a single trapped atom and a single photon. *Nature* **428**, 153–157 (2004).
57. Julsgaard, B., Sherson, J., Cirac, J. I., Fiurasek, J. & Polzik, E. S. Experimental demonstration of quantum memory for light. *Nature* **432**, 482–486 (2004).
58. Gottesman, D. & Chuang, I. L. Demonstrating the viability of universal quantum computation using teleportation and single-qubit operations. *Nature* **402**, 390–393 (1999).
59. Knill, E., Laflamme, R. & Milburn, G. A scheme for efficient quantum computation with linear optics. *Nature* **409**, 46–52 (2000).
60. Sanaka, K., Jennewein, T., Pan, J.-W., Resch, K. & Zeilinger, A. Experimental nonlinear sign shift for linear optical quantum computation. *Phys. Rev. Lett.* **92**, 017902 (2004).
61. Pittman, T. B., Fitch, M. J., Jacobs, B. C. & Franson, J. D. Experimental controlled-NOT logic gate for single photons in the coincidence basis. *Phys. Rev. A* **68**, 032316 (2003).
62. O'Brien, J. L., Pryde, G. J., White, A. G., Ralph, T. C. & Branning, D. Demonstration of an all-optical controlled-NOT gate. *Nature* **426**, 264–267 (2003).
63. Gasparoni, S., Pan, J.-W., Walther, P., Rudolph, T. & Zeilinger, A. Realization of a photonic controlled-NOT gate sufficient for quantum computation. *Phys. Rev. Lett.* **93**, 020504 (2004).
64. Raussendorf, R. & Briegel, H. J. A one-way quantum computer. *Phys. Rev. Lett.* **86**, 5188–5191 (2001).
65. Walther, P. et al. Experimental one-way quantum computing. *Nature* (in the press).
66. Steane, A. Multiple particle interference and quantum error correction. *Proc. R. Soc. Lond. A* **452**, 2551–2577 (1995).
67. Shor, P. in *Proc. 37th Symp. on Foundations of Computer Science* 15–65 (IEEE Computer Society Press, Los Alamitos, 1996).
68. Kuhn, A., Hennrich, M. & Rempe, G. Deterministic single-photon source for distributed quantum networking. *Phys. Rev. Lett.* **89**, 679011–679014 (2002).
69. McKeever, J. et al. Deterministic generation of single photons from one atom trapped in a cavity. *Science* **303**, 1992–1994 (2002).
70. Santori, C., Fattal, D., Vuckovic, J., Solomon, G. S. & Yamamoto, Y. Indistinguishable photons from a single-photon device. *Nature* **419**, 594–597 (2002).
71. Grangier, P., Sanders, B. & Vuckovic, J. (eds) Focus on Single Photons on Demand *New J. Phys.* (Spec. Edn) **6**, 85–100 (2004).
72. Brattke, S., Varcoe, B. T. H. & Walther, H. Generation of photon number states on demand via cavity quantum electrodynamics. *Phys. Rev. Lett.* **86**, 3534–3537 (1999).
73. Kim, J., Takeuchi, S., Yamamoto, Y. & Hogue, H. H. Multiphoton detection using visible light photon counter. *Appl. Phys. Lett.* **74**, 902–904 (1999).
74. Slusher, R. E., Hollberg, L. W., Yurke, B., Mertz, J. C. & Valley, J. F. Observation of squeezed states generated by four-wave mixing in an optical cavity. *Phys. Rev. Lett.* **55**, 2409–2412 (1985).
75. Furusawa, A. et al. Unconditional quantum teleportation. *Science* **282**, 706–709 (1998).
76. Schleich, W. P. *Quantum Optics in Phase Space* (Wiley-VCH, Berlin, 2001).
77. Borchers, H.-A. & Ralph, T. C. *A Guide to Experiments in Quantum Optics* 2nd edn (Wiley-VCH, Weinheim, 2004).
78. Grosshans, F., Assche, G. V., Wenger, J., Cerf, R. B. J. & Grangier, P. Quantum key distribution using gaussian-modulated coherent states. *Nature* **421**, 238–241 (2003).
79. Dirac, P. A. M. Emission and absorption of radiation. *Proc. R. Soc. Lond. A* **114**, 243–265 (1927).
80. Everett, H. “Relative state” formulation of quantum mechanics. *Rev. Mod. Phys.* **24**, 454–462 (1957).
81. Bohm, D. A suggested interpretation of the quantum theory in terms of “hidden” variables. *Phys. Rev.* **85**, 166–179, 180–193 (1952).
82. Rauschenbeutel, A. et al. Coherent operation of a tunable quantum phase gate in cavity QED. *Phys. Rev. Lett.* **83**, 5166–5169 (1999).
83. Jennewein, T., Simon, C., Weihs, G., Weinfurter, H. & Zeilinger, A. Quantum cryptography with entangled photons. *Phys. Rev. Lett.* **84**, 4729–4732 (2000).
84. Lenard, P. Über die lichtelektrische Wirkung. *Ann. Phys.* **8**, 149–198 (1902).
85. Millikan, R. A. A direct photoelectric determination of Planck's *h*. *Phys. Rev.* **7**, 355–388 (1915).
86. Millikan, R. A. Albert Einstein on his seventieth birthday. *Rev. Mod. Phys.* **21**, 343–345 (1949).
87. Compton, A. H. The spectrum of scattered X-rays. *Phys. Rev.* **22**, 409–413 (1923).
88. Bothe, W. & Geiger, H. Über das Wesen des Comptoneffekts; ein experimenteller Beitrag zur Theorie

- der Strahlung. *Z. Phys.* **32**, 639–663 (1925).
89. Stuewer, R. *The Compton Effect: Turning Point in Physics* (Science History, New York, 1975).
90. Tu, L.-C., Luo, J. & Gillies, G. T. The mass of the photon. *Rep. Prog. Phys.* **68**, 77–130 (2004).
91. Cocconi, G. Upper limits on the electric charge of the photon. *Am. J. Phys.* **60**, 750–751 (1992).
92. Weinberg, S. Light as a fundamental particle. *Phys. Today* **28**, 35–37 (1975).
93. Lawrence, E. O. & Beams, J. W. The element of time in the photoelectric effect. *Phys. Rev.* **32**, 478–485 (1928).
94. Forrester, A. T., Gudmundsen, R. A. & Johnson, P. Photoelectric mixing of incoherent light. *Phys. Rev.* **90**, 1691–1700 (1955).
95. Clauser, J. F. Experimental limitations to the validity of semiclassical radiation theories. *Phys. Rev. A* **6**, 49–54 (1972).
96. Klein, M. J. Einstein's first paper on quanta. *Nat. Phil.* **2**, 59–86 (1963).
97. Stachel, J. *et al.* in *The Collected Papers of Albert Einstein* (eds Stachel, J. *et al.*) Vol. 2, 134–148 (Princeton Univ. Press, Princeton, 1989).
98. Pais, A. Einstein and the quantum theory. *Rev. Mod. Phys.* **51**, 863–914 (1979).
99. Clauser, J. F. in *Quantum [Un]speakables: From Bell to Quantum Information* (eds Bertlmann, R. A. & Zeilinger, A.) 61–98 (Springer, Heidelberg, 2002).

Acknowledgements We acknowledge wonderful collaborations and challenging discussions with many colleagues and friends in the worldwide quantum optics community over the years.

Competing interests statement The authors declare that they have no competing financial interests.

In search of symmetry lost

Frank Wilczek

The Department of Physics, Massachusetts Institute of Technology, Cambridge, Massachusetts 02139, USA (e-mail: wilczek@mit.edu)

Powerful symmetry principles have guided physicists in their quest for nature's fundamental laws. The successful gauge theory of electroweak interactions postulates a more extensive symmetry for its equations than are manifest in the world. The discrepancy is ascribed to a pervasive symmetry-breaking field, which fills all space uniformly, rendering the Universe a sort of exotic superconductor. So far, the evidence for these bold ideas is indirect. But soon the theory will undergo a critical test depending on whether the quanta of this symmetry-breaking field, the so-called Higgs particles, are produced at the Large Hadron Collider (due to begin operation in 2007).

It has been almost four decades since our current, wonderfully successful theory of the electroweak interaction was formulated^{1–4}. Central to that theory is the concept of spontaneously broken gauge symmetry^{5,6}. According to this concept, the fundamental equations of physics have more symmetry than the actual physical world does. Although its specific use in electroweak theory involves exotic hypothetical substances and some sophisticated mathematics, the underlying theme of broken symmetry is quite old. It goes back at least to the dawn of modern physics, when Newton postulated that the basic laws of mechanics exhibit full symmetry in three dimensions of space — despite the fact that everyday experience clearly distinguishes ‘up and down’ from ‘sideways’ directions in our local environment. Newton, of course, traced this asymmetry to the influence of Earth’s gravity. In the framework of electroweak theory, modern physicists similarly postulate that the physical world is described by a solution wherein all space, throughout the currently observed Universe, is permeated by one or more (quantum) fields that spoil the full symmetry of the primary equations. Thus, modern physicists hypothesize that what we perceive as empty space is actually a highly structured medium. In fact, as I will elaborate below, we vehemently suspect that the world is a multilayered, multicoloured, cosmic superconductor.

Fortunately this hypothesis, which might at first hearing sound quite extravagant, has testable implications. The symmetry-breaking fields, when suitably excited, must bring forth characteristic particles: their quanta. Using the most economical implementation of the required symmetry breaking, one predicts the existence of a remarkable new particle, the so-called Higgs particle. More ambitious speculations suggest that there should be not just a single Higgs particle, but rather a complex of related particles. The very popular and attractive idea of low-energy supersymmetry^{7,8}, to be discussed further below, requires at least five ‘Higgs particles’.

The primary goal of fundamental physics is to discover profound concepts that illuminate our understanding of nature. Discovering new particles, as such, is secondary. In recent times, however, physicists have often found that their most profound concepts, when implemented with rigorous logic, are reflected in the existence of new particles. This happens because both quantum mechanics and special relativity are important in the regime of short distances and high energies, where high-energy physics explores fundamental laws. It is difficult to combine quantum mechanics and special relativity in a consistent way. The only way we know how to do it is by using quantum field theory, and the basic objects of quantum field theory are space-filling entities (quantum fields) whose excitations are what we perceive,

concretely, as particles^{9,10}. So, when our concepts are made consistent with quantum mechanics and relativity, they tend to be reflected rather directly in predictions about particles.

The *W* and *Z* bosons, carriers of the weak nuclear force, and gluons, carriers of the strong nuclear force, are outstanding examples of ideas embodied as particles. These so-called gauge particles are physical embodiments of the symmetry of physical law (gauge invariance)^{11–13} — not merely metaphorically but in a very precise sense. Indeed, as a fact of history, the existence of these particles and their detailed behaviour was predicted before their experimental observation, starting from the concept of gauge symmetry. Harmony between mind and matter, in the form of mathematical abstractions conjuring up sensuous reality, has long figured in the dreams of mystics and the inspiration of visionaries — the ‘music of the spheres’. Here it is realized in a form that is genuine, reproducible and precise.

Now we are faced with the opportunity for another synthesis. Ironically, the concept whose embodiment we now seek is a special, structured sort of symmetry breaking. This concept is a necessary complement to what has come before; for our symmetry-based understanding of the *W* and *Z* bosons — that is, of the electroweak interaction — relies on postulating symmetries that are broken in a very specific way. They are supposed to be spoiled by a form of cosmic superconductivity, with newly hypothesized fields having the role performed by electrons in ordinary superconductors. It is these new quantum fields that are the progenitors of Higgs particles.

So far, no Higgs particle has been observed. As yet, this failure does not represent a crisis. Detection of Higgs particles that are sufficiently heavy — specifically, those whose mass exceeds 114 GeV, which is the current lower bound¹⁴ — will have to await more powerful accelerators than are now available. But theory tells us that this evasion cannot be maintained indefinitely. If the Higgs particle, or an appropriate complex of Higgs particles, does not turn up at the Large Hadron Collider (LHC), a major revision of our thinking will be required. The LHC is now under construction at CERN (European Centre for Particle Physics) near Geneva. It is due to begin operation in late 2007.

There are already indirect but significant indications that at least one Higgs particle with a mass below 250 GeV does exist¹⁵. If there is such a particle, it will certainly be observed at the LHC. That observation, if and when it occurs, will bring a glorious chapter in physics to a glorious conclusion. It will also provide a key to unlock new volumes that are currently sealed; for the circle of ideas around symmetry breaking and the Higgs particle includes, quite close to its elegant central core, some of the darkest and most forbidding zones of ignorance in the existing landscape of

fundamental physics. That is exciting because it means we will have an opportunity to learn. Big ideas and speculations about the unification of forces, and the cosmology of the early Universe, as well as supersymmetry, are very much in play. Thus the Higgs sector is not only a destination, but also a portal.

There is a vast technical literature on most of the topics discussed here^{16,17} and quite a few popular and semipopular presentations. My goal here is to present a brief but substantial and critical review of the main concepts and prospects that is accessible to scientifically sophisticated non-experts, yet reflects the essence of present-day thinking.

The three standard systems

It has become conventional to say that our knowledge of fundamental physical law is summarized in a 'standard model'. But this convention lumps together three quite different conceptual structures. Instead, it is more accurate and informative to say that our current, working description of fundamental physics is made from three distinct parts: three 'standard systems'. These are the gauge system, the gravity system and the Higgs system.

Each of these systems concerns the interactions of a specific kind of particle: gauge bosons, gravitons and Higgs particles, respectively. It is remarkable that everything we know, or reliably infer, about the fundamental laws of nature can be interpreted as a statement about how one or another of these particles interacts with other forms of matter. To be precise, every known departure from trivial 'free' propagation — every nonlinear coupling of the quantum fields that describe matter in all its forms — involves interaction with a gauge particle, a graviton or a Higgs particle. Two of these three systems, the gauge and gravity systems, are governed by principled theories founded on deep, powerful concepts. Because of this they are tight, both conceptually and algorithmically.

The gauge system is constructed as the embodiment of extensive symmetries involving transformations among different kinds of 'colour' degrees of freedom. Colour used in this sense has nothing to do with optical phenomena; rather, it is a generalization of the concept of electromagnetic charge. Quantum chromodynamics (QCD) — our theory of the strong interaction^{18,19} — works with three colour charges. The weak interaction invokes two other colour charges and ordinary electromagnetism introduces yet another charge. The precise symmetry is expressed in the language of group theory, as $SU(3) \times SU(2) \times U(1)$.

Gauge symmetries, when combined with the principles of quantum mechanics and special relativity, are extremely powerful. Gauge symmetries require the existence of appropriate gauge bosons, and vice versa. Through this connection between mathematics and physics — concept and reality — we arrive at a beautiful and tightly integrated theory of gauge bosons and their interactions with other forms of matter. A profound reflection of this is that the physics of the gauge system is almost fully determined from considerations of inner consistency. In other words, it contains few freely adjustable parameters. There are basically just three such parameters, one each for $SU(3)$, $SU(2)$ and $U(1)$. Given these three numbers, no further 'fudge factors' remain available. The gauge system provides precise predictions for many phenomena; predictions that are in excellent agreement with numerous accurate experiments. In fact no significant deviation has been found so far. Each term in the foundational equations underlying a complete description of strong, weak and electromagnetic interactions has been checked out repeatedly in precise experiments.

The gravity system is essentially Einstein's general theory of relativity. It is sometimes claimed that general relativity is impossible to reconcile with quantum mechanics, that there is a terrible crisis in quantum gravity, and so on. On hearing this, one might be puzzled about how physicists and astrophysicists manage to get on with their work. The truth is that there is a very concrete and precise working theory of gravity that fully conforms to the principles of quantum mechanics and which so far has proved adequate to describe all physical and astrophysical observations. It is simply the quantum field

theory version of general relativity (for experts: the Einstein–Hilbert action for gravity itself, extended to matter using the minimal coupling procedure). This theory fails to give predictions for processes involving ultra-high-energy particles, but the energies at which it fails are much larger than those that are accessible to observation.

Like the gauge system, the gravity system is constructed as the embodiment of a powerful symmetry principle, in this case Einstein's general covariance. General covariance both requires the existence of the graviton, and tightly constrains its properties. It thereby generates a unique, principled theory of gravity. General relativity is fully specified in terms of just two freely adjustable parameters: one is Newton's gravitational constant G_N ; the other is the so-called cosmological term, which parameterizes the energy density of empty space. Until very recently there were only upper bounds on the value of the cosmological term, but observations of the acceleration of distant galaxies, together with indirect inference from measured cosmic microwave background anisotropies, seem to require a positive value^{20–22}. All other gravitational phenomena are predicted using only G_N . General relativity has scored many triumphs, both qualitative ones, such as providing foundations for Big Bang cosmology and black hole physics, and quantitative ones, such as accurately predicting the precession of Mercury and the time variation of binary pulsar frequencies.

The gauge system and the gravity system can both be written in appealing geometric forms. More precisely, quantum fields that describe different forms of curvature produce both gravitons and gauge bosons. For gravitons, the relevant curvature is that of space-time; for gauge bosons, it is curvature in so-called internal spaces, which are defined by the variables that describe configurations of colour charges. The coupling of these particles to other fields is also defined geometrically, technically, by the promotion of ordinary derivatives into covariant ones.

The third system, where the Higgs particle and its couplings to other forms of matter reside, is another story entirely. We know of no deep principle, comparable to gauge symmetry or general covariance, which constrains the values of these couplings at all tightly. In the Higgs system, the number of freely adjustable parameters mushrooms into dozens. Nor do the values of the couplings we now infer from the masses and mixing of quarks and leptons conform to any easily discernible pattern (see the section '...and a nest of kluges' below).

Clearly the Higgs system of fundamental physics is its least satisfactory part. Whether measured in terms of the large number of independent parameters it requires, or in terms of the small number of powerful ideas it contains, our theory of this sector falls far short of the high level we have achieved elsewhere. In particular, despite the phrase's connotation, no 'theory of everything' hitherto proposed has in practice materially improved our theory of the Higgs system. Having placed the Higgs system in context, let us now scrutinize it more closely.

Symmetry breaking in superconductivity

The most fundamental phenomenon of superconductivity is the Meissner effect, according to which magnetic fields are expelled from the bulk of a superconductor. The Meissner effect implies the possibility of persistent currents²³. Indeed, if a superconducting sample is subjected to an external magnetic field, currents of this sort must arise near the surface of a sample to generate a cancelling field.

An unusual but valid way of speaking about the phenomenon of superconductivity is to say that within a superconductor the photon acquires a mass. The Meissner effect follows from this. Indeed, to say that the photon acquires a mass is to say that the electromagnetic field becomes a massive field. Because the energetic cost of supporting massive fields over an extended volume is prohibitive, a superconducting material finds ways to expel magnetic fields.

Bardeen, Cooper and Schrieffer (BCS) developed a satisfactory microscopic theory of superconductivity in metals²⁴. BCS theory traces superconductivity to the existence of a special sort of long-range correlation among electrons. This effect is purely quantum-

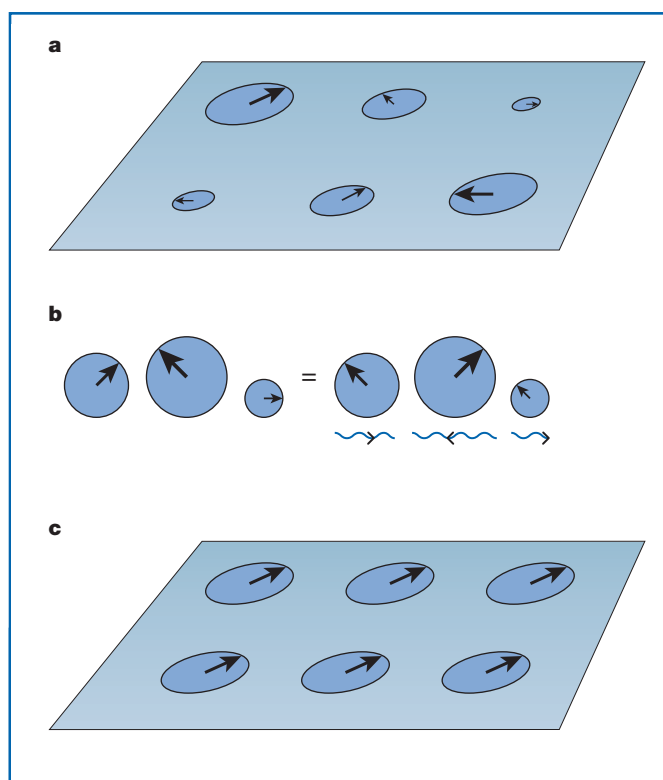


Figure 1 Visual metaphors for gauge symmetry and superconductivity. **a**, The quantum-mechanical wavefunction of an electrically charged field is a position-dependent complex number, which can be depicted as a collection of arrows with different sizes and orientations. **b**, Gauge symmetry states that the same physical situation can be described using different orientations of the arrows (phases of the wavefunction) with a compensating gauge field (electromagnetic potential, represented by wavy arrows). This is somewhat analogous to Einstein's equivalence principle, by which relatively accelerated reference frames become equivalent when a compensating gravity field is included. **c**, In the superconducting state, the charged-particle wavefunction is correlated over long distances and times — whereas in the non-superconducting state (**a** could be regarded as a typical, disordered slice of this state), the wavefunctions fluctuate randomly in space and time. The ordered structure of **c** is disrupted by gauge transformations, so now the electromagnetic potentials have an energetic cost — the photon has acquired a mass.

mechanical. A classical phenomenon that is only very roughly analogous, but much simpler to visualize, is the occurrence of ferromagnetism owing to long-range correlations among electron spins (that is, their mutual alignment in a single direction). The sort of correlations responsible for superconductivity are of a much less familiar sort, as they involve not the spins of the electrons, but rather the phases of their quantum-mechanical wavefunctions. One cannot do full justice to this concept without some elaborate mathematical preparation. But as it is the leading idea guiding our construction of the Higgs system, I think it is appropriate to sketch an intermediate picture that is more accurate than the magnet analogy and suggestive of the generalization required in the Higgs system.

First we imagine that the electrons organize themselves into pairs, the so-called Cooper pairs. The wavefunction of a Cooper pair is a complex-valued function; it has both an amplitude and a phase. If we have a uniform density of Cooper pairs, then the amplitude is constant, but the phase can vary in space and time. We can represent the different possible values of the phase by points on a circle. So in representing the quantum dynamics of the Cooper pairs at each point of space-time, we have an overlying circular 'internal space' — an extra dimension if you like — and the position of the wavefunction in this extra dimension must be specified (Fig. 1a).

A fundamental principle of electrodynamics is its gauge symmetry. A gauge transformation is a mathematical transformation of electromagnetic potentials and the wavefunctions of charged particles. When we say electrodynamics obeys gauge symmetry, we mean that while the gauge transformation changes the variables that appear in the equations of electrodynamics, and consequently rearranges those equations, it nevertheless leaves the overall physical content of the equations unchanged (Fig. 1b). Weyl and London discovered the gauge symmetry of quantum electrodynamics in the 1920s (refs 11, 12). Yang and Mills proposed a more general concept of gauge symmetry¹³ that supports a much wider class of transformations (non-abelian gauge theory). At first this generalization seemed to be a mathematical curiosity, but as physics developed it has come to be more and more central. We have come to recognize both that gauge symmetry is necessary for the consistency of the theory, and conversely, that the equations of electrodynamics can be derived from gauge symmetry, assuming the validity of quantum mechanics and special relativity. One aspect of this is that gauge symmetry enforces zero mass for the photon.

For present purposes, what is crucial is that gauge transformations rotate the wavefunction in the extra dimension, through an angle that can vary depending on location in space-time. Superconductivity occurs when the phases of the Cooper pairs all align in the same direction; that is, when they all have the same position within their extradimensional circles (Fig. 1c). Of course, gauge transformations that act differently at different space-time points will spoil this alignment. Thus, although the basic equations of electrodynamics are unchanged by gauge transformations, the state of a superconductor does change. To describe this situation, we say that in a superconductor gauge symmetry is spontaneously broken.

The phase alignment of the Cooper pairs gives them a form of rigidity. Electromagnetic fields, which would tend to disturb this alignment, are rejected. This is the microscopic explanation of the Meissner effect, or in other words, the mass of photons in superconductors.

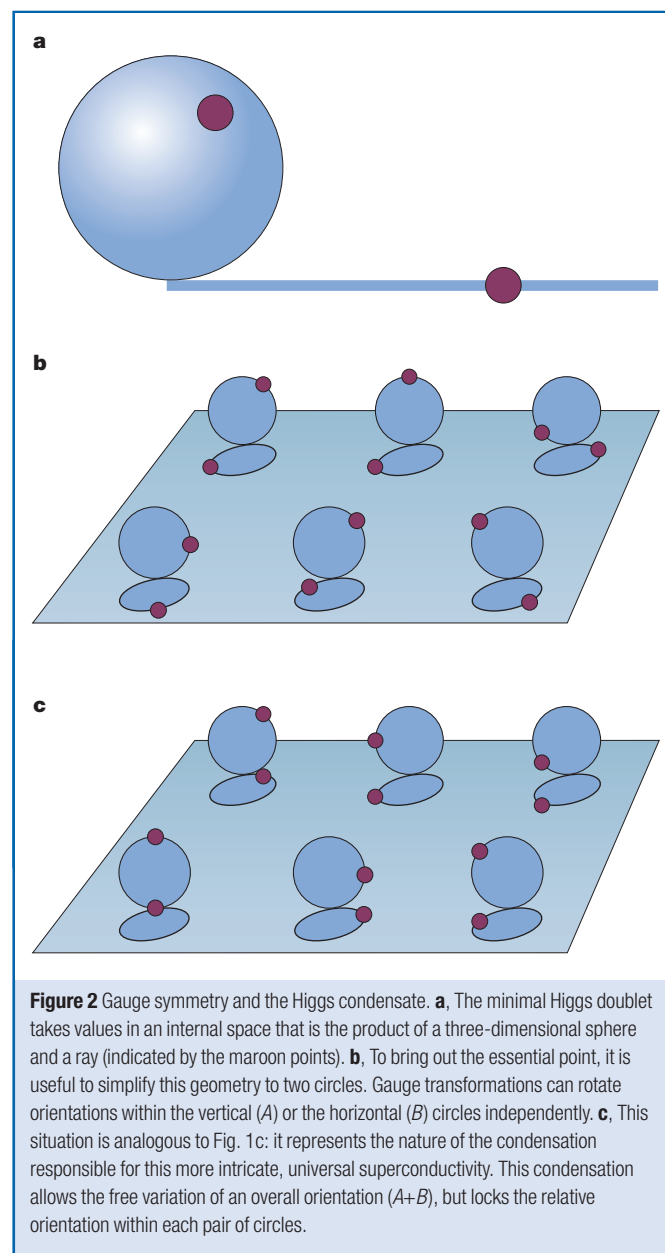
Electroweak symmetry breaking: some elegant details...

Several basic properties of the $W^\pm$ and Z particles, which are responsible for mediating the weak interaction, are quite similar to properties of photons. They are all spin-1 particles that couple with universal strength to appropriate charges and currents. This resemblance, together with consistency requirements of the sort mentioned above, suggests that the equations of the weak interactions must have an appropriate gauge symmetry, distinct from, but of the same general nature as, the gauge symmetry of electromagnetism.

Elaboration of this line of thought leads to our modern theory of the electroweak interactions, which is firmly based on the postulate of gauge symmetry. The gauge symmetry involved is more elaborate than that of electromagnetism. It involves a more intricate internal space (Fig. 2a). Instead of just a circle (orientation) and a ray (magnitude), we have a three-dimensional sphere (orientation) and a ray (magnitude) over each point of space-time, so four 'internal dimensions'. Gauge transformations rotate the spheres, although not all types of rotation are allowed. Mathematically, the allowed transformations define the group $SU(2) \times U(1)_Y$.

There is an essential supplement to this generalization, which provides our primary motivation for postulating the existence of the Higgs system. Unlike the photon, which acquires mass only inside superconductors, $W^\pm$ and Z are massive particles even in empty space. The equations of gauge symmetry, in their pristine form, predict the existence of these particles, but require that their masses vanish. To account for the masses of $W^\pm$ and Z , we must suppose that what we perceive as empty space is, in reality, a new form of superconductor, not for electromagnetism, but for its near-relation gauge interactions.

With this interpretation, what we perceive as empty space is not so empty. But what is it that has the role — for this new universal superconductivity — of the Cooper pairs whose alignment is responsible for conventional superconductivity? No form of matter identified so



far provides a suitable candidate. We are therefore led to postulate that there is a new form of matter doing the job. Accordingly, what we perceive as empty space is in fact filled with an exotic, suitably aligned substrate: the Higgs condensate.

‘Observing’ this condensate only through its effect on *W* and *Z* bosons gives us very limited insight into its nature. The minimal hypothesis, within the conventional framework of contemporary physics, is to postulate the existence of a quantum field with just enough structure to support the desired pattern of symmetry breaking, imparting mass to some, but not all, of the gauge fields.

Let me spell this out with reference to the slightly simplified model of Fig. 2b. The pattern we want in reality is $SU(2) \times U(1)_Y \rightarrow U(1)_\gamma$; the simplified model exhibits $U(1)_A \times U(1)_B \rightarrow U(1)_{A+B}$, but the essential points are the same.

We begin with two independent rotation symmetries, but only a combination survives, which will represent ordinary electromagnetism. To break the symmetry in the appropriate way, we require the condensate to have a component in both circles (Fig. 2c). Furthermore, its orientations within the circles must be equal to one another, although their common angle can vary. Similarly to what was discussed for the Cooper-pair condensate, the Higgs condensate has a

form of rigidity that imposes an energetic price on fields that attempt to disrupt this favourable pattern. But the field that rotates both circles by the same amount preserves the pattern, so it remains massless. It can be identified with the electromagnetic field, which produces the photon γ . Note that the $A+B$ gauge symmetry of electromagnetism ($U(1)_\gamma$) is not the same as either ‘pure’ rotation *A* or *B* of the primary circles.

In the realistic $SU(2) \times U(1)_Y \rightarrow U(1)_\gamma$ case, a similar entwining of weak and electromagnetic gauge symmetry is mandatory, simply because the $W^\pm$ bosons are electrically charged. Thus, these interactions must be treated together, and we speak of ‘electroweak’ theory. It is often said that the $SU(2) \times U(1)_Y \rightarrow U(1)_\gamma$ theory of broken gauge symmetry is a unified theory of weak and electromagnetic interactions, but that seems to me an overstatement. The theory still involves two quite separate primary symmetries, which are related only by the fact that the same Higgs condensate spoils them both. The simplest way of implementing the symmetry-breaking pattern $SU(2) \times U(1)_Y \rightarrow U(1)_\gamma$ makes specific predictions for the masses and couplings of the *W* and *Z* bosons, which seem to be quite accurate. More complicated implementations are conceivable, but they seem neither desirable nor necessary.

Together with its orientations in the internal sphere, the quantum field associated with the Higgs condensate must have another degree of freedom, representing its overall magnitude. The degrees of freedom corresponding to variations in orientation are associated with the (broken-symmetry) gauge transformations. They get absorbed into the massive gauge fields; their quanta are longitudinally polarized *W* and *Z* bosons. The degree of freedom associated with changes in overall magnitude, however, has independent meaning. Its elementary excitation, or quantum, is what is usually referred to as the Higgs particle. The inescapable minimal consequence of our postulate of a Higgs condensate is a new, electrically neutral spin-0 particle.

One embellishment of the minimal scheme is especially well motivated, because it arises in connection with the important concept of low-energy supersymmetry²⁵. The generalization simply involves imagining that two independent fields of the same general character (both representing positions on internal spheres and circles, and undergoing the same sorts of gauge transformations) contribute to the Higgs condensate. A striking consequence is that we then expect to have not one but five new Higgs particles, because we have introduced a field with four additional degrees of freedom (specified by its location within its four-dimensional internal space). Of these, three are electrically neutral; the other two are positively and negatively charged, each being the other’s antiparticle.

...and a nest of kluges

It is not only the masses of *W* and *Z* bosons that are inconsistent with pristine $SU(2) \times U(1)$ gauge symmetry and that get tied up with the Higgs condensate. The masses of quarks and leptons present a similar difficulty. In the end, these masses too can be ascribed to interaction with the Higgs condensate, but the details are quite different and frankly (in my view) seem quite ugly and clearly provisional.

The proximate source of the difficulty is the observed parity violation of the weak interaction²⁶. In its most basic form, the observed phenomenon is that when fast-moving quarks or leptons are emitted in weak decays, their spin tends to be aligned opposite to their direction of motion; they are therefore said to be left-handed. To accommodate this in the framework of gauge symmetry, we would like to say that the symmetry acts only on left-handed particles. That formulation, however, is inconsistent with special relativity. Indeed, a sufficiently fast-moving observer could overtake the particle, and to such an observer, its direction of motion would appear reversed, and it would be right-handed. This difficulty would be avoided if the quarks and leptons had zero mass, for then they would move at the speed of light and could not be overtaken. So we can implement the interactions of *W* bosons with

massless left-handed quarks and leptons in a consistent way using gauge symmetry; but non-zero masses for quarks and leptons must be tied up with gauge symmetry breaking.

Unlike in the case of W and Z bosons, there are no nice geometrical pictures associated with the masses of quarks and leptons, nor connections with profound concepts (gauge invariance, superconductivity). We are reduced to simply writing down the equations. For non-expert readers, the details are not vital, but I do want to convey a sense of what leads me to call them 'ugly' and 'provisional'.

At present there is no compelling theory that predicts the values of any quark or lepton masses. Yet there are striking facts to be explained that play a crucial part in the structure of the world as we know it. The mass of the electron, and of the up and down quarks are nearly a million times smaller than the 'natural' scale (250 GeV) set by the magnitude of the Higgs condensate, but the mass of the top quark is close to that scale; the CKM matrix (that describes the mixing of quark species) is measured to be almost, but not quite, the unit matrix. All this presumably indicates that principles — or perhaps accidents or even conspiracies — are lurking within the Higgs sector, which ensure that some, but not all, of the possible couplings are small. It is a great challenge for the future to discover these principles, or alternatively to understand why accident and conspiracy run rampant. So far, we have only had access to the masses and the CKM matrix, which encode highly processed versions of the basic couplings. The discovery of Higgs particles and measurements of their interactions will give us better access to fundamentals.

On the origin(s) of mass

As we have just seen, the masses of W and Z bosons, and of quarks and leptons, arise from the interaction of these particles with the pervasive Higgs condensate. This has inspired references to the Higgs particle as 'the origin of mass', or even 'the God particle'²⁷. The real situation is interesting but rather different from what this hyperbole suggests. A few critical comments seem in order.

First, most of the mass of ordinary matter has an entirely different origin. This mass is contained in atomic nuclei, which are built up from nucleons (protons and neutrons), which in turn are built up from quarks (mainly up and down quarks) and colour gluons. Colour gluons are strictly massless, and the up and down quarks have tiny masses, compared with the mass of nucleons. Instead, most of the mass of nucleons (more than 90%) arises from the energy associated with the motion of the quarks and gluons that compose them²⁸ — according to the original form of Einstein's famous equation, $m = E/c^2$. This circle of ideas provides an extraordinarily beautiful, overwhelmingly positive answer to the question Einstein posed in the title of his original paper²⁹: "Does the inertia of a body depend on its energy content?" And it has nothing to do with Higgs particles!

Second, as we have just seen, for quarks and leptons the Higgs mechanism appears more as an accommodation of mass than an explanation of its origin. We map observed values of masses and mixings through some distorting lenses into corresponding values of Higgs-field couplings, but only for the W and Z bosons do we have reliable, independent insight into what these values ought to be. And third, the Higgs field in no sense explains the origin of its own mass. A parameter directly equivalent to that mass must be introduced into the equations explicitly.

Finally, there is no necessary connection between mass and interaction with any particular Higgs field. As an important example, much of the mass of the Universe is observed to be in some exotic, 'dark' form. The dark matter has only been observed indirectly, through the influence of its gravity on surrounding ordinary matter. It has evaded more conventional observation using optical, radio or other telescopes, so evidently it has only a feeble coupling to photons. We do not yet know what this dark stuff is precisely, but the leading theoretical candidates (axions and weakly interacting massive particles, WIMPs) are massive particles that do not acquire their mass by interacting with the electroweak Higgs condensate. The general

point is that many kinds of hypothetical particle can have masses that, unlike the masses of W and Z bosons and quarks and leptons, do not violate the electroweak $SU(2) \times U(1)$ symmetry, and such masses need have no relation to that symmetry's spoiler, the electroweak Higgs condensate. The Higgs particle's own mass is also of this kind.

The genuine meaning of the Higgs field — that it embodies the concept of symmetry breaking and makes tangible the vision of a universal cosmic superconductor — is deep, strange, glorious, experimentally consequential and very probably true. This meaning has no need for embellishment, and can only be diminished by dubious oversimplification. So blush for shame, ye purveyors of hyperbole! End of sermon.

Searching for Higgs

From the observed masses of the W and Z bosons we infer the magnitude of the Higgs condensate to be 250 GeV. This sets the scale for structure in the Higgs system. The mass of the lightest excitation of the condensate could only be significantly larger than this if a large parameter appeared in the theory to amplify the scale. More specifically, what is relevant to this question is the coefficient of the nonlinear self-coupling of the Higgs field. But large self-coupling seems unlikely on both theoretical and experimental grounds. Theoretically, it takes us into a regime where the Higgs field undergoes violent quantum fluctuations, and appears to develop inconsistencies³⁰. On the experimental side, precision measurements of various quantities can be compared to theoretical predictions at a level that is sensitive to the contribution of 'virtual particles'; that is, quantum fluctuations, including contributions that arise from coupling to the Higgs system. All indications so far are that these contributions are small, consistent with weak coupling¹⁵. Thus, on very general grounds, we expect to find new particles of some kind, recognizable as quanta of the Higgs condensate, with mass most probably smaller than 250 GeV, and in any case not much larger.

To go further and be quantitative, we must consider more specific models. The minimal implementation of symmetry breaking, which

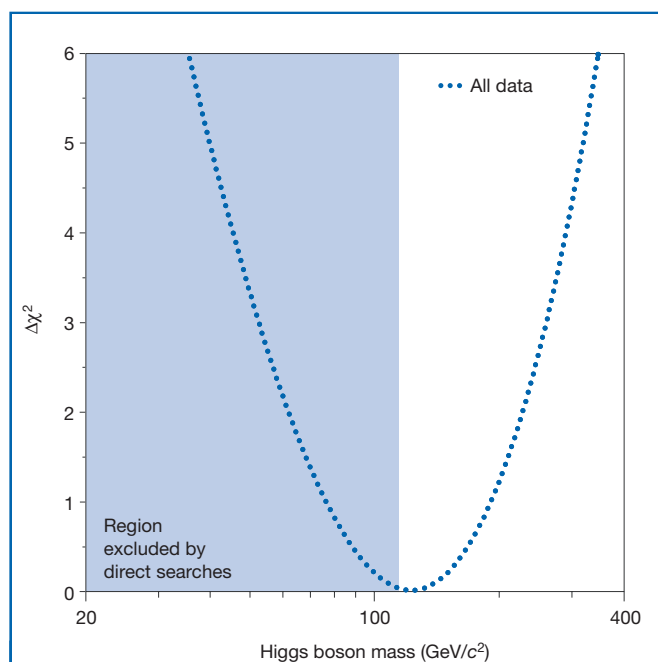


Figure 3 Constraints on the mass of the Higgs boson from experimental data. The dotted line shows, as a function of the Higgs mass, the χ^2 for a global fit to electroweak data (including such quantities as the mass of the W boson and the Z boson total width). Also included in the dataset is the current world-average value for the mass of the top quark⁴⁸. A relatively light value for the Higgs mass is favoured, less than 200 GeV. Shading indicates the region excluded by direct searches for the Higgs boson¹⁴.

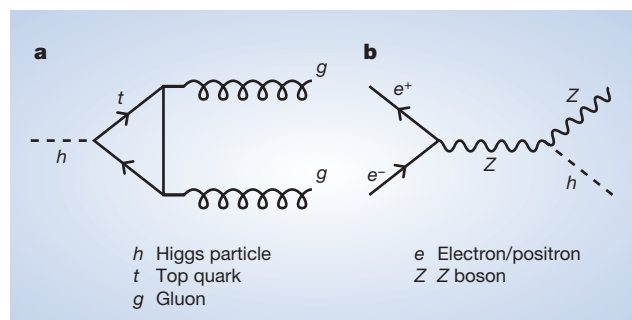


Figure 4 Higgs couplings. **a**, The dominant coupling of the Higgs particle to ordinary matter arises indirectly, through a virtual top/anti-top quark pair that annihilates into gluons. Direct couplings of the Higgs to up or down quarks, electrons, gluons or photons are very small, due to the tiny masses of these particles. **b**, Electrons can, however, give access to the Higgs through their coupling to the Z boson. The *Bremsstrahlung* process shown produces a distinctive Zh final state. Similar processes involving quark/antiquark annihilation produce either Zh or Wh final states.

assumes that only a single $SU(2) \times U(1)$ doublet exists (and thereby, as discussed above, predicts the existence of exactly one physical Higgs particle), is of course the canonical choice. Within this framework, which I shall adopt for the next few paragraphs, everything of interest is determined in terms of just one unknown parameter, conveniently taken to be the value of the Higgs particle's mass.

Now, with everything else pinned down, the aforementioned contributions (due to virtual particles, and often called radiative corrections) to a variety of measured quantities can be calculated precisely, as a function of the Higgs particle's mass. When this is done, one finds that the central value of the Higgs mass preferred by the experimental results is about 100 GeV, with a dispersion of about 50 GeV (Fig. 3).

The strength with which the Higgs particle couples to other fundamental particles is proportional to the mass of the particle involved (with a most important exception that I will come to in a moment). Thus, the dominant decay mode of the Higgs particle will be into the heaviest available particles. For a wide range of masses, including the most favoured values, the dominant decay is therefore into a heavy quark pair — bottom/anti-bottom ($b\bar{b}$); channels with heavier particles are energetically forbidden. The rates of decay into other quark or lepton pairs — charm/anti-charm and tau/anti-tau — are lower but not insignificant.

The exceptional case is gluons³¹. The direct coupling of the Higgs particle to gluons, inferred from the classical form of the theory, vanishes — as one might expect from the vanishing mass of the gluons. There is a purely quantum-mechanical process, however, in which the Higgs particle fluctuates into a virtual top/anti-top quark pair, which then materializes into a real gluon pair (Fig. 4a). As a fraction of all possible decays, this branch amounts to about 10%. A coupling to photons is similarly achieved by fluctuations through virtual W and Z bosons, as well as top quarks; it is smaller, owing to the weaker photon coupling.

Search strategies for the Higgs particle at electron–positron accelerators (the Large Electron Positron collider, LEP, or the planned International Linear Collider, ILC) and at proton–proton accelerators (the Tevatron at Fermilab, United States, or CERN's LHC) are quite different. The coupling of electrons and positrons to the Higgs particle is extremely small. A favoured strategy exploits processes in which the electron and positron annihilate into a Z boson, which then radiates the Higgs particle³² (Fig. 4b). This has the further advantage that the final-state Z boson is very well characterized and can be reliably identified. Indeed, a general problem in Higgs-hunting is that one must distinguish the process of interest

from possible 'background' processes that can lead, from the point of view of what is observable in practice, to effectively the same final state. This channel (with a final state of $Zb\bar{b}$) was the subject of very intense scrutiny at LEP, especially during the final stages of its operation in 2000. The result, however, was inconclusive. A lower bound of 114.1 GeV, at the 95% confidence level, was put on the mass of the Higgs particle (assuming the minimal implementation). On the other hand, statistically weak (1.7σ) hints of excess events were observed at the upper limit of the machine's energy, consistent with what would be induced by a Higgs particle of mass 116 GeV.

At a hadron machine of sufficiently large energy, such as the LHC, Higgs particles are produced copiously through the gluon-fusion process, which exploits the Higgs' indirect coupling to gluon pairs³³. The difficulty here is not production, but the signal-to-noise ratio in detection. Final states at hadron machines typically contain much extraneous debris, because the initial projectiles — protons — are complex objects. Indeed, our gluon-fusion production process relies on the fact that protons, in addition to quarks, contain a substantial fraction of gluons. The dominant final result of Higgs decay, a $b\bar{b}$ pair, is also very easily produced by ordinary particle interactions that in no way involve the Higgs particle. These form, from an experimental point of view, an almost impenetrable background. For this reason, it might be best to focus on the rarer decay of the Higgs into a pair of photons. Another strategy is to look for 'tagged' events that contain a W or Z in the final state, accompanying the Higgs particle (as at electron–positron machines). These could result from the radiation of Higgs particles from the W or Z bosons produced by quark/antiquark annihilation.

Once we depart from the minimal implementation, many more possibilities for probing the Higgs sector open up, some much less demanding. For example, the additional Higgs particles predicted in the minimal supersymmetric model include electrically charged ones. These have very characteristic decay modes and they could be copiously produced at electron–positron machines once the energetic threshold for pair production is passed.

At this point it is appropriate to mention that whereas in the minimal model the predicted value of the Higgs mass is only weakly constrained from above, the minimal supersymmetric model generically predicts the existence of at least one Higgs particle with mass below 140 GeV, whose properties resemble those of the (absolutely) minimal Higgs particle with the same mass^{17,34}. The above-mentioned indications from radiative corrections, and perhaps from LEP, might be interpreted as encouraging for the supersymmetric alternative.

This account of the practical strategy of Higgs searches is no more than a sketch of some of the simplest considerations involved. A vast, multi-faceted and growing literature is devoted to the subject. There is a general consensus that discovery of one or more Higgs particles cannot elude experimenters at the LHC, but that an exploration of the Higgs system worthy of the opportunities it affords will require the clean environment and high energies available in electron–positron collisions at the ILC.

Extensions: unification, supersymmetry

If we merely count particles, then the minimal implementation of gauge symmetry breaking — using one doublet field and leading to one physical Higgs particle — recommends itself on grounds of economy. But considerations of logical coherence and structural integrity seem to lead us in another direction.

The structure of the gauge system gives powerful suggestions for its further fruitful development³⁵. The product structure $SU(3) \times SU(2) \times U(1)$, the reducibility of the fermion representation (that is, the fact that the symmetry does not make connections linking all the fermions), and the peculiar values of the quantum number hypercharge assigned to the known particles all suggest a larger symmetry. The devil is in the details and it is not at all automatic that the observed, complex pattern of matter will fit neatly into a simple mathematical structure. But, to a remarkable extent, it does (Box 1).

Box 1

Unification

Within the standard gauge system the strong interactions are described by quantum chromodynamics (QCD), a theory based on the gauge symmetry group $SU(3)$ defined by transformations among

a

$$\begin{pmatrix} u & u & u \\ d & d & d \end{pmatrix}_{1/6} \quad \begin{pmatrix} u^c & u^c & u^c \end{pmatrix}_{-2/3} \quad \begin{pmatrix} d^c & d^c & d^c \end{pmatrix}_{1/3} \quad \begin{pmatrix} \nu \\ e \end{pmatrix}_{-1/2} \quad e^c_1$$

b

Quantum number Particle	<i>R</i>	<i>W</i>	<i>B</i>	<i>G</i>	<i>P</i>
<i>u</i>	+	−	−	+	−
<i>u</i>	−	+	−	+	−
<i>u</i>	−	−	+	+	−
<i>d</i>	+	−	−	−	+
<i>d</i>	−	+	−	−	+
<i>d</i>	−	−	+	−	+
<i>u</i> ^c	−	+	+	−	−
<i>u</i> ^c	+	−	+	−	−
<i>u</i> ^c	+	+	−	−	−
<i>d</i> ^c	−	+	+	+	+
<i>d</i> ^c	+	−	+	+	+
<i>d</i> ^c	+	+	−	+	+
<i>ν</i>	+	+	+	+	−
<i>e</i>	+	+	+	−	+
<i>e</i> ^c	−	−	−	+	+
<i>N</i>	−	−	−	−	−

$$\text{Hypercharge } Y = -\frac{1}{6}(R + W + B) + \frac{1}{4}(G + P)$$

three colour charges; the weak interactions are described by an independent but mathematically similar gauge symmetry $SU(2)$ using two colour charges, and finally an independent ‘hypercharge’ symmetry $U(1)$ based on a single type of charge. Ordinary electromagnetism involves a combination of $SU(2)$ and $U(1)$ symmetry that remains valid after these separate symmetries are broken by the Higgs condensate.

In part **a** of the figure, the relationships of the quarks and leptons under these transformations are shown. The strong gauge symmetries act horizontally, the weak gauge symmetries act vertically and the values of the hypercharges are indicated by subscripts. The 15 quarks and leptons shown here fall into five unrelated clans. (There is a three-fold repetition of this entire structure, accommodating 45 quarks and leptons altogether.)

Theories of unified gauge symmetry propose that there is a more extensive symmetry that involves transformations among all these colour charges. That symmetry must be spontaneously broken, to explain why we observe the consequences of different types of charge to be quite different (strong interactions really are strong, and weak interactions weak). Nevertheless, this idea is not without consequences: the quarks and leptons must furnish the material for building unified structures that remain coherent under the extended symmetry.

One particular unified symmetry passes this test with flying colours (and is shown in part **b** of the figure). Although the smallest simple group into which $SU(3) \times SU(2) \times U(1)$ could possibly fit is $SU(5)$ (ref. 49) — it fits all the fermions of a single family into two representations (**10** + **5**) and the hypercharges click into place — a larger symmetry group, $SO(10)$ (ref. 50), fits these and one additional $SU(3) \times SU(2) \times U(1)$ singlet particle into a single representation (the spinor **16**). All 15 quarks and leptons appear on the same footing, and the additional particle, which has the quantum numbers of a right-handed neutrino, is quite welcome: it plays a crucial role in the attractive ‘seesaw’ model for neutrino masses^{43,44}.

Where before we had a piecemeal accommodation of the observed particles, now we have a marvellous correspondence between reality and a unique, ideal mathematical object.

This unification of quantum numbers, although attractive, remains purely formal until it is embedded in a physical model. This requires realizing the enhanced symmetry in a local gauge theory. The enhanced symmetry must be broken. The Higgs system of electroweak theory supplies a precedent for that. What we need is another condensate, with a vastly larger density. We are proposing that the world is a multi-coloured, multi-layered superconductor.

There is an apparent difficulty with these ideas, which on closer scrutiny turns out to represent their greatest success. Non-abelian gauge symmetry requires that the relative strengths of the different couplings must be equal (universality), which is not what is observed. Fortunately, there is a compelling way to save the situation³⁶. The higher symmetry is broken at a very large energy scale (equivalently, a small distance scale), but we observe interactions at much smaller energies (larger distances). The strength we observe differs from the intrinsic strength, because it is affected by the physics of vacuum polarization. A cloud of virtual particles surrounds the charge and can enhance or dilute its power. The resulting ‘running’ of couplings is an effect that can be calculated quite precisely, given a definite hypothesis about the particle spectrum. In this way we can test quantitatively the idea that the observed couplings originate from a single unified value at small distances.

The results of these calculations are quite remarkable and encouraging (Fig. 5). If we include vacuum polarization from the particles we know about in the minimal standard model, we find approximate

unification. If we include vacuum polarization from the particles needed in expanding the standard model to include supersymmetry^{37,38}, softly broken at the TeV (1,000 GeV) scale, we find accurate unification^{39,40}. The unification occurs at a very large energy scale, of order 10^{16} GeV. This success is robust against small changes in the supersymmetry-breaking scale, and is not adversely affected by incorporation of additional particle multiplets, as long as they form complete representations of $SU(5)$ (ref. 37).

Low-energy supersymmetry is desirable on several other grounds as well. The most important has to do with the Higgs condensate. In the absence of supersymmetry, radiative corrections to the magnitude of this condensate diverge (‘radiative corrections’ is standard jargon for the effect of virtual particles or, alternatively, quantum fluctuations). One must fix the condensate’s value, which sets the scale for electroweak symmetry breaking, by hand, as part of the definition of the theory. This ‘renormalization’ procedure leaves it utterly mysterious why the empirical value is so much smaller than unification scales. Moreover, enhanced unification symmetry requires that the Higgs doublet should come together with additional fields, to fill out a complete representation. Its partners, however, have the quantum numbers to mediate proton decay, so if they exist at all, their masses must be very large, at about the unification scale of 10^{16} GeV. This reinforces the idea that such a large mass is what is ‘natural’ for a scalar field⁴¹. The relatively small mass (of the order of 100 GeV) of the Higgs field that we need in our electroweak theory

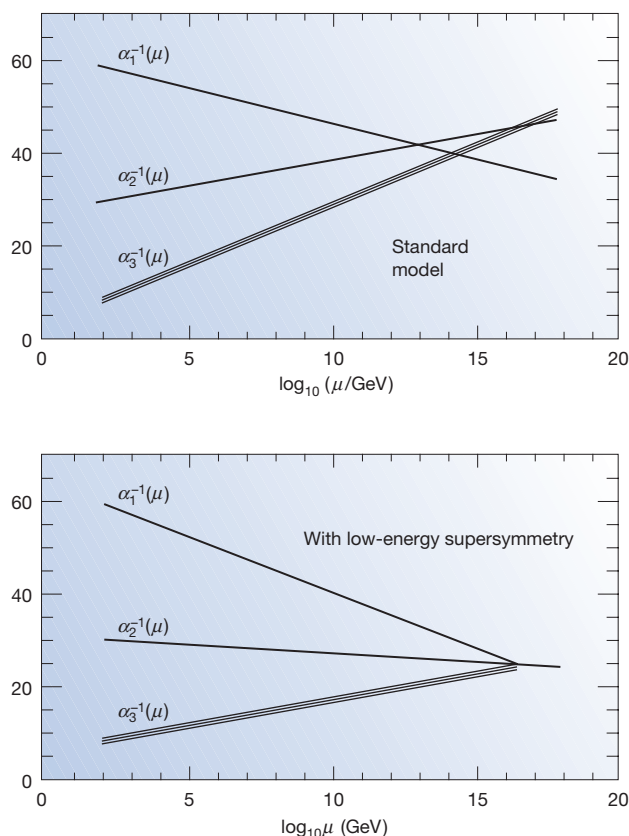


Figure 5 Unification of the forces. The strengths of the couplings of the weak, electromagnetic and strong forces are hugely disparate (represented here as α_1^{-1} , α_2^{-1} and α_3^{-1}). But their perceived strength changes with the energy scale of the process (μ), through corrections due to virtual particles. Assuming there are only the particles known to us in the standard model and extrapolating beyond the reach of experiment to very high energies, the couplings move towards each other but do not converge at a single point (top). If, however, the extra particles needed to implement low-energy supersymmetry are included in the calculation, the couplings meet neatly at an energy of about 10^{16} GeV (lower plot). Note that the energy scale is logarithmic (and the existence of other unknown particles is overlooked), so this calculation is a bold — perhaps reckless — extrapolation of the laws we know to apply to energies vastly larger than those at which these laws have been tested.

seems unnatural and requires some special justification. Supersymmetry, if it is not too badly broken, largely solves this problem, for it ensures that these unsavoury radiative corrections are small⁴².

The fact that our unification calculations point to an enormous new mass scale for unification is profound. This enormous mass scale is inferred entirely from data taken at much lower energies. The disparity of scales arises from the slow (logarithmic) running of inverse couplings, which implies that modest differences in observed couplings must be made up by a long interval of running. The appearance of a very large mass scale is welcome on several grounds. I will mention three of the most important.

First, right-handed neutrinos, which as we have seen can enhance the symmetry of unification, naturally acquire masses of the order of the unification scale. Masses of that magnitude remove these particles from direct experimental accessibility, but they can have a most important indirect effect^{43,44}. This is because, in second-order perturbation theory, the ordinary left-handed neutrinos make virtual transitions to their right-handed relatives and back. This exotic process generates non-zero masses for the ordinary neutrinos, but these are much smaller than the masses of other leptons and quarks.

The magnitudes that arise in this way are broadly consistent with the tiny observed masses of neutrinos. No more than order-of-magnitude success can be claimed because many relevant details of the models are poorly determined.

Second, unification tends to obliterate the distinction between quarks and leptons, and hence to open up the possibility that protons decay (their building-block quarks turn into electrons or muons). Heroic experiments to observe this process have so far come up empty-handed, with limits on partial lifetimes approaching 10^{34} years for some channels. It is very difficult to ensure that these processes are sufficiently suppressed, unless the unification scale is very large. Even the high scale indicated by the running of couplings and neutrino masses is barely adequate. Spinning it positively, experiments to search for proton decay remain a most important and promising probe into the physics of unification. Similarly, it is difficult to avoid the idea that unification brings in new connections among the different families. There are significant experimental constraints on flavour-changing neutral currents, lepton number violation and other exotic processes that must be suppressed, and this makes a high mass scale for the virtual particles that mediate them most welcome.

Third, with the appearance of this large scale, unification of the strong and electroweak interactions with gravity becomes much more plausible. Newton's constant has dimensions of mass, so it runs even classically. Or, to put it another way, gravity responds to energy/momentum, so it gets stronger at large energy scales. Nevertheless, because gravity starts out extremely feeble compared to other interactions on laboratory scales, it becomes roughly equipotent with them only at enormously high scales, comparable to the Planck energy of 10^{18} GeV. By inverting this thought, we gain a deep insight into one of the main riddles about gravity: if gravity is a primary feature of nature, reflecting the basic structure of space-time, why does it ordinarily appear so feeble? Elsewhere⁴⁵, I have traced the answer down to the fact that, at the unification (Planck) scale, the strong coupling is about $1/2$!

In view of all this, our accounting of the 'economy of ideas' is altered. For it seems that with five Higgs particles you can buy a lot more than with one.

Cosmological implications

In the very early Universe, when temperatures were much higher, the Higgs condensate that now fills all space could not have maintained its alignment over extended distances. In a word, it melted⁴⁶. Just as a superconductor heated beyond its critical temperature goes normal, or a magnet heated above its Curie temperature loses its magnetization, the Universe would then have been in a different, more symmetric phase. In this phase, W and Z bosons — like photons, colour gluons and gravitons — had zero mass, as did quarks and leptons. (Ironically, the Higgs particles themselves retained a finite mass.)

Thus, during the early evolution of the Universe there was a dramatic change in the properties of matter. The detailed physical nature of this change is at present unknown. It may have been a sharp phase transition in the thermodynamic sense, or a smooth crossover. Such a cosmic phase transition might have been accompanied by unusual or violent physical events that left lasting consequences. One possibility is that the current imbalance between the abundance of matter and antimatter might have been generated when the Higgs condensate froze in. Another is that the Higgs freeze-in catalysed an epoch of extremely rapid cosmic acceleration, akin to or even identical to the inflationary epoch, whose occurrence is widely conjectured in modern cosmology⁴⁷ but whose physical nature is highly uncertain. It is only by studying the Higgs system in detail that we can begin to assess these possibilities reliably.

The existence of a Higgs system with properties of the general sort I have discussed, notably including one or more accessible, recognizable Higgs particle, appears to be a compelling consequence of quantum field theory and the standard model of fundamental physics

—a complex of ideas that has been tremendously successful. This would be a beautiful thing to observe, and extremely instructive.

And yet, our standard system of gravity — general relativity — incorporates the principle that all forms of energy exert gravitational influence. As a special case, the postulated Higgs condensate, which fills all space, should weigh something. In fact, it should weigh a lot: estimating the energy density of this condensate using straightforward dimensional reasoning gives a value much larger than is allowed by observations. It would show up as such a large contribution to the cosmological term that the size of the Universe would double every 10^{-38} seconds! Thus, either the Higgs condensate does not exist, or its energy density is cancelled out by some other, still more exotic, contribution, or there is a profound lacuna in our understanding of gravity. The second and third alternatives seem sufficiently improbable as to suggest that just maybe we will be dragged to the first. If so, then our search for Higgs particles as the quanta of gauge symmetry breaking might instead turn up something quite different. Theoretical physicists, roused from their dogmatic slumbers, would be forced back to the drawing board. Wouldn't that be fun? □

doi:10.1038/nature03281

1. Glashow, S. Partial symmetries of weak interactions. *Nucl. Phys.* **22**, 579–588 (1961).
2. Weinberg, S. A model of leptons. *Phys. Rev. Lett.* **19**, 1264–1266 (1967).
3. Salam, A. in *Elementary Particle Physics, Nobel Symp.* (ed. Svarthom, N.) No. 8, 367–377 (Almqvist & Wiksell, Stockholm, 1968).
4. 't Hooft, G. & Veltman, M. Regularization and renormalization of gauge fields. *Nucl. Phys. B* **44**, 189–213 (1972).
5. Englert, F. & Brout, R. Broken symmetry and the mass of gauge vector mesons. *Phys. Rev. Lett.* **13**, 321–323 (1964).
6. Higgs, P. Broken symmetries and the masses of gauge bosons. *Phys. Rev. Lett.* **13**, 508–509 (1964).
7. Wess, J. & Zumino, B. A lagrangian model invariant under supergauge transformations. *Phys. Lett. B* **49**, 52–54 (1974).
8. Ferrara, S. (ed.) *Supersymmetry* (North Holland/World Scientific, Singapore, 1987).
9. Weinberg, S. *The Quantum Theory of Fields 1: Foundations* (Cambridge Univ. Press, Cambridge, 1995).
10. Wilczek, F. Quantum field theory. *Rev. Mod. Phys.* **71**, S85–S95 (1999).
11. Weyl, H. Z. Elektron and gravitation I. *Z. Phys.* **56**, 330–352 (1929).
12. London, F. Quantenmechanische deutung der theorie von Weyl. *Z. Phys.* **42**, 375–389 (1927).
13. Yang, C. N. & Mills, R. Conservation of isotopic spin and isotopic gauge invariance. *Phys. Rev.* **96**, 191–195 (1954).
14. ALEPH Collaboration, DELPHI Collaboration, L3 Collaboration, OPAL Collaboration & The LEP Working Group for Higgs Boson. Search for the standard model Higgs boson at LEP. *Phys. Lett. B* **565**, 61–75 (2003).
15. LEP Electroweak Working Group [online] <http://lepewwg.web.cern.ch/LEPEWWG> (2004)
16. Einhorn, M. (ed) *The Standard Model Higgs Boson* (North Holland, Amsterdam, 1991).
17. Gunion, J., Haber, H., Kane, G., & Dawson, S. *The Higgs Hunter's Guide* (Addison Wesley, New York, 1990).
18. Gross, D. & Wilczek, F. Ultraviolet behavior of non-Abelian gauge theories. *Phys. Rev. Lett.* **30**, 1343–1346 (1973).
19. Politzer, H. Reliable perturbative results for strong interactions? *Phys. Rev. Lett.* **30**, 1346–1349 (1973).
20. Riess, A. *et al.*, Observational evidence from supernovae for an accelerating universe and a cosmological constant. *Astron. J.* **116**, 1009–1038 (1998).
21. Perlmutter, S. *et al.* Measurements of Ω and Λ from 42 high-redshift supernovae. *Astrophys. J.* **517**, 565–586 (1999).
22. Spergel, D. *et al.*, Determination of cosmological parameters. *Astrophys. J.* **148** (suppl.), 175–194 (2003).
23. London, F. *Superfluids 1: Macroscopic Theory of Superconductivity* (Wiley, New York, 1950).
24. Bardeen, J., Cooper, L. & Schrieffer, R. Theory of superconductivity. *Phys. Rev.* **108**, 1175–1204 (1957).
25. Weinberg, S. *Quantum Theory of Fields*, Vol. 3: *Supersymmetry* (Cambridge Univ. Press, Cambridge 2000).
26. Lee, T. D. & Yang, C. N. Question of parity conservation in weak interactions. *Phys. Rev.* **104**, 254–258 (1956).
27. Lederman, L. & Teresi, D. *The God Particle* (Bantam, New York, 1993).
28. Wilczek, F. The origin of mass. *Ann. Phys. (MIT)* **16**, 24–35 (2003).
29. Einstein, A. Does the inertia of a body depend upon its energy content? [in German] *Ann. Phys.* **18**, 639–641 (1905).
30. Kuti, J., Lin, L. & Shen, Y. Upper bound on the Higgs-boson mass in the standard model. *Phys. Rev. Lett.* **61**, 678–681 (1988).
31. Wilczek, F. Decays of heavy vector mesons into Higgs particles. *Phys. Rev. Lett.* **39**, 1304–1306 (1977).
32. Bjorken, J. in *Proc. SLAC Summer School* (SLAC publication 198, 1976).
33. Georgi, H., Glashow, S., Machacek, M. & Nanopoulos, D. Higgs bosons from two-gluon annihilation in proton–proton collisions. *Phys. Rev. Lett.* **40**, 692–694 (1978).
34. CMS collaboration. The compact muon solenoid [online] <http://www.lip.pt/~outreach/docs/cms2> (2004)
35. Pati, J. & Salam, A. Unified lepton–hadron symmetry and a gauge theory of the basic interactions. *Phys. Rev. D* **8**, 1240–1251 (1973).
36. Georgi, H., Quinn, H. & Weinberg, S. Hierarchy of interactions in unified gauge theories. *Phys. Rev. Lett.* **33**, 451–454 (1974).
37. Dimopoulos, S., Raby, S. & Wilczek, F. Supersymmetry and the scale of unification. *Phys. Rev. D* **24**, 1681–1683 (1981).
38. Dimopoulos, S. & Georgi, H. Softly broken supersymmetry and SU(5). *Nucl. Phys.* **193**, 150–162 (1981).
39. Amaldi, U., de Boer, W. & Fürstenau, H. Comparison of grand unified theories with electroweak and strong coupling constants measured at LEP. *Phys. Lett. B* **260**, 447–455 (1991).
40. Ellis, J., Kelly, S. & Nanopoulos, D. Precision LEP data, supersymmetric GUTs and string unification. *Phys. Lett. B* **249**, 441–448 (1990).
41. Gildener, E. & Weinberg, S. Symmetry breaking and scalar bosons. *Phys. Rev. D* **13**, 3333–3341 (1976).
42. Witten, E. Dynamical breaking of supersymmetry. *Nucl. Phys. B* **188**, 513–554 (1981).
43. Gell-Mann, M., Ramond, P. & Slansky, R. in *Supergravity: Proceedings of the Supergravity Workshop at Stony Brook, September 27–29, 1979* (eds van Nieuwenhuizen, P. & Freedman, D.) 315–321 (North Holland, Amsterdam, 1979).
44. Yanagida, T. in *Workshop on Unified Theory and Baryon Number in the Universe* (eds Sawada, O. & Sugamoto, A.) 95–98 (KEK, Tsukuba, 1979).
45. Wilczek, F. Scaling mount Planck 3: is that all there is? *Phys. Today* **55N8**, 10–11 (2002).
46. Kirzhnits, D. & Linde, A. Symmetry behaviour in gauge theories. *Ann. Phys.* **101**, 195–238 (1976).
47. Guth, A. Inflationary universe: a possible solution to the horizon and flatness problems. *Phys. Rev. D* **23**, 347–356 (1981).
48. A precision measurement of the mass of the top quark. *Nature* **429**, 638–642 (2004).
49. Georgi, H. & Glashow, S. Unity of all elementary-particle forces. *Phys. Rev. Lett.* **32**, 438–441 (1974).
50. Georgi, H. in *Particles and Fields 1974* (ed. Carlson, C.) 575–584 (AIP, New York, 1975).

Competing interests statement The author declares that he has no competing financial interests.

The state of the Universe

Peter Coles

School of Physics & Astronomy, University of Nottingham, University Park, Nottingham NG7 2RD, UK (e-mail: Peter.Coles@nottingham.ac.uk)

The past 20 years have seen dramatic advances in cosmology, mostly driven by observations from new telescopes and detectors. These instruments have allowed astronomers to map out the large-scale structure of the Universe and probe the very early stages of its evolution. We seem to have established the basic parameters describing the behaviour of our expanding Universe, thereby putting cosmology on a firm empirical footing. But the emerging 'standard' model leaves many details of galaxy formation still to be worked out, and new ideas are emerging that challenge the theoretical framework on which the structure of the Big Bang is based. There is still a great deal left to explore in cosmology.

The idea of the Big Bang has been with us for more than half a century. During that time, its character has changed dramatically in several ways. Until the early years of the twentieth century, cosmology was primarily a metaphysical subject. With the arrival of Einstein's general theory of relativity¹, a sensible mathematical description^{2,3} of the space-time structure of the entire Universe emerged in the 1920s (Box 1). Spurred on by Hubble's discovery of the cosmic expansion⁴, theorists in the 1930s began to speculate about how the Universe got to be in its present state, what physical processes might have been involved in its evolution, and what it might have been like in the past. These lines of enquiry gradually brought new areas of science, such as nuclear and particle physics, into play. The resulting confrontation between theoretical physics and observational astronomy, within a context provided by Einstein's theory, has been essential to progress in this field. The latest phase of this story has been dominated by stunning observational developments that have strengthened a growing belief that answers have been obtained to many of the basic questions that cosmologists have been asking for decades. But this is nowhere near the end of the road for cosmology. Many important issues within the standard model remain to be resolved, especially in the area of galaxy formation. Moreover, the newest physics of string theory has thrown up a number of theoretical ideas that may initiate a revolution in our understanding of the early Universe and overthrow some of the central concepts of the Big Bang.

The Friedmann models (Box 1) and their associated interpretational framework have allowed the construction of an astonishingly successful broad-brush description of the evolution of the Universe that accounts for most available observational data. Besides the Hubble expansion⁴, the main evidence in favour of the Big Bang theory has been the discovery⁵, by Penzias and Wilson in 1965, of the cosmic microwave background (CMB) radiation⁶. Although the spectrum of this radiation was not well determined at first, the COBE satellite⁷ revealed an astonishingly accurate black-body behaviour with a temperature around 2.73 K. Using the Friedmann framework allows us to turn the clock back to an epoch when the Universe was about one-thousandth of its present size, at which point the radiation temperature was a thousand times higher. This would be hot enough to ionize atomic matter, which would then have scattered radiation well enough to be effectively opaque. Although it is very difficult to imagine how such an accurate black-body spectrum could have been made naturally at the low temperature at which it exists now, dense plasmas can maintain sufficient thermal contact with radiation to establish a thermal spectrum. The microwave background was

last scattered by such material about 300,000 years after the Big Bang. Turning the clock back further, to an epoch less than a minute after the Big Bang, the matter's temperature is measured in billions of degrees. At this time, the Universe resembles a thermonuclear explosion during which protons and neutrons are rapidly baked into light nuclei in a heat bath provided by the background radiation. Gamow⁸, and Alpher and Herman⁹ showed, for example, that in such conditions helium can be created much more easily than is accomplished by stellar hydrogen burning and with a much lower contamination of heavier elements. Deuterium, helium-3 and lithium-7 are also made in trace quantities during the Big Bang.

The production of the CMB and the synthesis of helium take place during conditions that are achievable in laboratory experiments. To go further into the early Universe, one needs to adopt more speculative theoretical considerations. When the Universe was around a microsecond old is the time when the hadrons split up into their constituent quarks and gluons. At about 10^{-11} seconds, the distinction between electromagnetic and weak nuclear reactions disappears with the restoration of electroweak symmetry, which is broken at low energies. Earlier still, the matter contents of the Universe are expected to be described by more speculative particle physics, such as grand unified theories (GUTs), so cosmologists sought ways of incorporating ideas from quantum field theory into the framework furnished by general relativity. These forays into particle cosmology have scored numerous successes, such as the idea of supersymmetric dark matter and — perhaps the most influential theoretical idea^{10–13} of the past 25 years — cosmic inflation (Box 2). Successful though inflation has been, it does represent a kind of half-way house on the road to a fully consistent description of the very early Universe. The marriage of low-energy effective quantum field theory with classical general relativity is an uncomfortable one, and in any case no fundamental scalar has yet been identified as a compelling candidate for the inflaton field or the form of its interaction potential.

More recent attempts to fuse quantum theory with gravity have focused on string theory, which is framed in spaces of higher dimensionality than the four used in the standard framework. The extra dimensions may be wrapped up on a very small scale so as to be unobservable, an old idea dating back to Kaluza¹⁴ and Klein¹⁵. On the other hand, in the latest 'braneworld' models, the extra dimensions can be large but we are confined to a four-dimensional subset of them (a 'brane'). In one version of this idea, dubbed the 'ekpyrotic universe'^{16,17}, the origin of our observable Universe lies in the collision between two branes floating around in a higher-dimensional 'bulk'. Other models are possible that result in the modification

Box 1

Big Bang basics

The theoretical structure of modern cosmology is based on a family of mathematical models derived from Einstein's general theory of relativity¹ in the 1920s by Friedmann² and Lemaitre³. Given the conceptual complexity of the underlying physics, it is surprising how simple these models are. In essence, they contain three components: (1) a description of the space-time geometry; (2) a set of equations describing the action of gravity; and (3) a description of the bulk properties of matter and energy.

To make progress with Einstein's theory, it is necessary to make some simplifying assumptions about the symmetry of the system to which it is applied. In the case where the system is the entire Universe this need is especially pronounced, and the remedy particularly drastic. The cosmological principle (CP) is the assumption that, on large scales, the Universe is homogeneous and isotropic. The adoption of the CP makes the description of space-time geometry pertaining to cosmological models extremely simple.

Space-time

Einstein's theory involves the use of a metric tensor g that relates four-dimensional space-time intervals to a general set of coordinates. The CP, however, furnishes a preferred time coordinate: any observer can define a 'clock' in terms of the local density of matter at their location. This defines cosmological time, t . All observers using this clock see the same matter density at any particular time. This simplifies the four-dimensional machinery of Einstein's theory into a much simpler '3+1' structure, and removes much of the complexity that arises where no special choice of time coordinate is obvious. Space-times compatible with the CP must have the same geometry at each point on a surface of constant time. The space-time may be expanding or contracting, however, so different time slices can differ by a scale factor $a(t)$. The possibilities are described by the Robertson–Walker metric:

$$ds^2 = c^2 dt^2 - a^2(t) \left[\frac{dr^2}{1 - kr^2} + r^2 (d\theta^2 + \sin^2\theta d\phi^2) \right] \quad (3)$$

in polar coordinates centred on the observer. There are only three possibilities: $k = 0$ represents a flat universe with a euclidean geometry on each time slice; $k > 0$ signifies a closed universe, with positively curved spatial surfaces like three-dimensional versions of the surface of a sphere; and $k < 0$ indicates negatively curved space sections of hyperbolic form. Light rays travel through this space-time along trajectories defined by $ds^2 = 0$.

Gravity

Einstein's field equations can be written in the form:

$$G_{\mu\nu} = \frac{8\pi G}{c^4} T_{\mu\nu} \quad (4)$$

The tensors $G_{\mu\nu}$ and $T_{\mu\nu}$ represent the other two components of the theory: $G_{\mu\nu}$ is the Einstein tensor, which describes the action of gravity through the curvature of space-time (this contains derivatives of the

metric g with respect to space-time coordinates); $T_{\mu\nu}$ is the energy-momentum tensor, which describes the bulk properties of matter. Here again the CP comes to our aid, as it forces the matter contents to have the form of a perfect fluid with some pressure p and energy density ρc^2 . The Einstein equations then simplify to the following:

$$3 \left(\frac{\dot{a}}{a} \right)^2 = 8\pi G\rho - \frac{3kc^2}{a^2} + \Lambda c^2 \quad (5)$$

$$\left(\frac{\ddot{a}}{a} \right) = -\frac{4\pi G}{3} \left(\rho + \frac{3p}{c^2} \right) + \frac{\Lambda c^2}{3} \quad (6)$$

$$\dot{\rho} = -3 \left(\frac{\dot{a}}{a} \right) \left(\rho + \frac{p}{c^2} \right) \quad (7)$$

Equations (5)–(7) determine the time evolution of the cosmic scale factor $a(t)$, and therefore describe the global expansion or contraction of the Universe; the overdots denote derivatives with respect to cosmic time.

Matter-energy

To solve this system of equations, we need to specify the equation of state that characterizes the material contents of the Universe. Cold (non-relativistic matter) can be described by a 'dust' equation of state with $p = 0$. If the Universe is filled with relativistic particles (either photons or very hot matter), then the appropriate equation of state is of the form $p = \rho c^2/3$. In the basic Big Bang theory, the early Universe is radiation-dominated; as it expands and cools, the matter becomes non-relativistic and the equation of state changes smoothly to that of dust.

The quantity Λ is the cosmological constant, introduced by Einstein as a modification of his original form of general relativity⁹⁷. His intention was to rescue what he thought of as a severe problem with his version of gravity theory. In the absence of the Λ -terms in equations (5) and (6), it is not possible to obtain a static solution obeying the CP. One can make the right-hand-side of equation (5) zero by balancing the terms in ρ and k . Having done that, however, one cannot also make the right-hand side of equation (6) zero. With a Λ -term, this difficulty is removed. Einstein thought of this as a modification of the law of gravity, and he subsequently referred to this as his 'greatest blunder'. His blunder was not so much the introduction of the cosmological constant, but rather his failure to realize that cosmological models obeying the CP should be dynamical. In other words, he failed to predict the Hubble expansion⁴.

Einstein could equally well have put the term on the other side of the field equations, in which case it would have been regarded as part of the description of the energy contents of the Universe. Viewed in this light, the cosmological constant (if it really is constant; that is, if it does not evolve with cosmic time) has exactly the same effect as a perfect fluid with an equation of state of the form $p = -\rho c^2$, as is clear from equation (7), which shows that the energy density is constant in this case.

of the Friedmann equations at early times¹⁸. These steps towards string cosmology are controversial, but they do motivate a radically different interpretation of early Universe physics. In these models, the low-energy Universe is described by an effective quantum field theory with the configuration of the vacuum state being controlled by the extra dimensions. In these theories there is a myriad of possible vacua, each with its own set of fundamental constants, different gauge hierarchies, and so on. It is possible that these ideas may also lead to cosmological models that avoid the initial singularity that occurs in the standard Big Bang, but this also remains an open question.

One of the concrete things that inflation has achieved is to bring us closer to an explanation of how galaxies and large-scale structure came into being. So far we have focused on a broad-brush description of the cosmos in terms of models obeying the cosmological principle. In reality, our Universe is not at all smooth and featureless. It contains stars, galaxies, clusters and superclusters of galaxies, and a vast amount of complexity. The basic idea for the origin of cosmic structure is simple. Whereas a perfectly smooth, self-gravitating fluid will remain homogeneous for ever, any irregularities — however slight — will be amplified by the action of gravity. This process is exponentially fast if the starting configuration is static¹⁹, but in an expanding universe²⁰ it must

Box 2

Cosmic inflation

The first fully-formulated version of cosmic inflation was produced by Guth¹⁰, although some of its components had been advanced by Starobinsky a couple of years earlier²³. In Guth's idea, there is a stage in the evolution of the Universe during which its energy density is dominated not by matter or relativistic particles but by the action of a scalar quantum field, the 'inflaton', ϕ . Such a field does not possess an equation-of-state of the type discussed in Box 1, but is instead characterized by an effective energy density ρc^2 and pressure p which depend on the form of its interaction potential, $V(\phi)$:

$$\rho c^2 = \frac{1}{2} \dot{\phi}^2 + V(\phi) \quad (8)$$

$$p = \frac{1}{2} \dot{\phi}^2 - V(\phi) \quad (9)$$

Guth realised that if a situation could be engineered in which the kinetic terms (involving derivatives of ϕ) could be made small compared to the potential terms, one could obtain an equation of state of the form $p = -\rho c^2$, just like the vacuum energy described in Box 1. A possible way of doing this is by allowing the Universe to undergo a phase transition in which the field rolls down the potential

into a new vacuum state. If this happens, the Universe can be made to accelerate, at least while the field remains in transit between states. This would cause the scale factor to increase exponentially for a very short time, driving the curvature term in the Friedmann equation (5) to zero and rendering the spatial surfaces effectively flat — however curved they were before the onset of inflation. After this extravagant burst of hyper-expansion, the Universe would be expected to revert to the more sedate radiation-dominated form. This basic concept has led to a plethora of variations: old inflation, new inflation, chaotic inflation, extended inflation, and so on^{11–13}.

Inflation and its descendants constitute a paradigm, rather than a theory. Nobody has yet identified the scalar field responsible. It is also not entirely clear what inflation actually predicts. It does predict a drastic reduction in the spatial curvature during the inflationary epoch, but there are initial conditions that still lead to appreciable curvature today. Its predictions therefore rely on some (often unstated) assumptions about the natural measure on the space of possible Big Bang universes. Nevertheless, it has furnished a coherent framework within which to trace the evolution of the early Universe and to make contact with observations, which is why it has played such an important role in the development of the subject.

overcome the expansion in order to initiate collapse. This means that cosmological structure formation is a relatively difficult process, which requires significant initial fluctuations to get it going. In inflationary models, quantum fluctuations in the scalar field driving the expansion generate adiabatic density perturbations with particularly simple statistical properties^{21,22} that 'seed' cosmic structure formation. They also generate an as-yet undetected background of primordial gravitational waves²³. Whereas the early stages of structure formation seem to be comprehensible, the hydrodynamical and radiative processes by which collapsed lumps turn into stars and galaxies are so complicated that even the sustained efforts of the most powerful supercomputers have left many questions open.

Cosmology by numbers

The Big Bang 'theory' is seriously incomplete. For the usual dust or radiation equations of state, the Universe is always decelerating. As it is expanding now, it must have been expanding more quickly in the past. There is therefore a stage, at a finite time in the past, at which the scale factor must shrink to zero and the energy density becomes infinite. This is the Big Bang singularity, at which the whole framework falls apart. We therefore have no way of setting the initial conditions that would allow us to obtain a unique solution to the system of equations (4)–(6) in Box 1. There is an infinitely large family of possible universes, so to identify which (if any) is correct we have to use observations rather than pure reason.

The simplest way to see how this is done is to re-write the Friedmann equation (equation (5)) in dimensionless form:

$$\Omega_m + \Omega_\Lambda + \Omega_k = 1 \quad (1)$$

The different terms in this equation relate to the different terms in equation (5). The ordinary matter density is expressed via $\Omega_m = 8\pi G\rho/3H^2$, the vacuum energy is $\Omega_\Lambda = \Lambda c^2/3H^2$ and $\Omega_k = -kc^2/a^2H^2$. (Here G is the gravitational constant, c is the velocity of light, Λ is the cosmological constant, and H is the expansion rate given in terms of the scale factor a by $H = (1/a)(da/dt)$; the density ρ , curvature k and scale factor a are defined in Box 1.). Note that Hubble's constant is involved in these definitions, so in what follows I will take $H_0 = 100h \text{ km s}^{-1} \text{ Mpc}^{-1}$ so that h is also dimensionless. These dimensionless parameters can in principle (and, now, in practice)

be measured. Relativistic particles (principally photons and light neutrinos) also contribute to equation (1), but their effect is negligible at late times. Cosmological nucleosynthesis provides a stringent constraint on the contribution of baryonic material (protons and neutrons), Ω_b , to the overall density of non-relativistic matter, Ω_m . For example, a high baryon abundance would result in more helium than observed, and vice versa for a low abundance. The value of $\Omega_b h^2 \approx 0.02$ is required to match the observations.

In 1994, G. Ellis and I wrote a review²⁴ in which we focused on all the available evidence about Ω_m . It is possible to estimate the mean density of the Universe (which is basically what this parameter measures) in many ways, including galaxy and cluster dynamics, galaxy clustering, large-scale galaxy motions, gravitational lensing, and so on. At the time, many of these 'weighing' techniques were in their infancy but, on the basis of what seemed to be the most robust evidence, we concluded that the evidence favoured a value $0.2 < \Omega_m < 0.4$. This conclusion still requires that most of the matter in the Universe be non-baryonic, perhaps in the form of some relic of the early Universe produced at such high energies that it has been impossible to identify in accelerator experiments. But whatever it is, there is not enough of it to close the Universe (that is, to make $\Omega_m = 1$). Our conclusion was somewhat controversial at the time: most theorists, but not all²⁵, then preferred a universe with $\Omega_m = 1$ because inflation suggested that the spatial sections should be flat (or very close to it; Box 2). In the absence of a cosmological constant, or dark energy, $\Omega_m = 1$ is needed to make a flat universe. Our conclusion about the value of Ω_m remains valid, but the past 10 years have seen two spectacular advances that we did not foresee at all, because there was no direct evidence at the time.

In our 1994 paper²⁴, we devoted only a very small space to the field of 'classical cosmology'. The idea of this approach is to use observations of distant objects (such as sources seen at appreciable redshifts) to directly probe the expansion rate and geometry of the Universe. For example, owing to the focussing effect of spatial curvature, a rod of fixed physical length would subtend a smaller angle when seen through an open universe (with $k < 0$) than in a closed universe (with $k > 0$; see Box 1). A standard light source would likewise appear fainter in a universe that is undergoing accelerated expansion than in a decelerating universe. The difficulty is that standard sources are difficult to come by. The Big Bang universe is an evolving system, so that

sources seen at high redshift are probes of an earlier cosmic epoch. Young galaxies are probably very different to mature ones, so they cannot be used as standard sources. Classical cosmology consequently fell into disrepute until, just a few years ago, it underwent a spectacular renaissance.

Dark energy

Two major programmes, the Supernova Cosmology Project²⁶ and The High-*z* Supernova Search²⁷, exploited the behaviour of a particular kind of exploding star, type Ia supernovae, as 'standard candles'. The special thing about this kind of supernova is that it is thought to result from the thermonuclear detonation of a carbon-oxygen white dwarf. These events are themselves roughly standard explosions, because the mass involved is always close to the Chandrasekhar mass, but they are not perfectly regular. The breakthrough was the discovery of an empirical correlation between the peak luminosity of the event and the temporal evolution of its subsequent fading. This correlation can be used to reduce the scatter from event to event. The results from both teams seem very conclusive. The measurements indicate that high-redshift supernovae are indeed systematically fainter than one would expect on the basis of extrapolation from similar low-redshift sources using decelerating world models. The results are sensitive to a complicated combination of Ω_m and Ω_Λ through the integral along the line of sight $\int_0^z \frac{dz'}{H(z')}$ from the source to the observer through the expanding space-time, as described in Box 1. These observations strongly favour accelerating world models. This in turn, strongly suggests the presence of a non-zero vacuum energy. More recently, evidence has emerged that the highest-redshift supernovae are probing an epoch during which the Universe was decelerating, showing that the onset of cosmic acceleration was relatively recent^{28,29} (Fig. 1). This is what one would expect, as the cosmological constant is constant, but the contribution of matter and radiation to the total energy budget should have been higher in the past.

In 1994 there was no direct evidence either for flat space or for Λ . One major concern, which at the time was clouded by large observational uncertainties, was whether the age of the oldest stars (in globular clusters) was consistent with the age of the Universe based on the cosmic dynamics (Box 1) and the measured value of the Hubble constant. This was a major problem for cosmologies with $\Omega_m = 1$, because such models have ages of $6.5h^{-1}$ Gyr, requiring $h = 0.5$ to be compatible even with a conservative lower limit on the ages of globular cluster stars. The evidence for Λ from supernova observations has also relaxed the pressure on this issue; accelerating world models are older than decelerating ones for the same value of h . Refinements of the cosmic distance scale³⁰ and globular cluster ages³¹ have subsequently brought observation and theory into good agreement, and provided further evidence for the Big Bang framework.

In the language of the Friedmann models, accelerating universes require the existence of a cosmological constant, as the density and pressure terms on the right-hand side of equation (6) are positive for 'normal' forms of matter. These terms arise because Einstein modified the left-hand side of equation (4) by the simple addition of a term involving Λ . Like the zero-point vacuum fluctuations discovered by Casimir³², it was realized that one-loop quantum fluctuations in 'empty' space should behave exactly in this way, so the natural description of the cosmological constant was as 'vacuum energy' or 'dark energy'; that is, as a sort of quantum correction to the energy-momentum tensor.

This exciting connection between microscopic quantum physics and large-scale cosmic dynamics is one of the biggest unexplained mysteries in modern science³³. The problem is that the vacuum energy is divergent, so should utterly dominate the curvature and expansion rate of the Universe. If the divergences are cut off at the Planck energy (hc^5/G)^{1/2} $\approx 10^{19}$ GeV, the resulting energy density is 123 orders of magnitude larger than observations allow. A cut-off at the quantum chromodynamics scale would result in a discrepancy of 40 orders of magnitude, which is still a long way off. It seems reasonable to suppose that such a drastic problem has a very fundamental

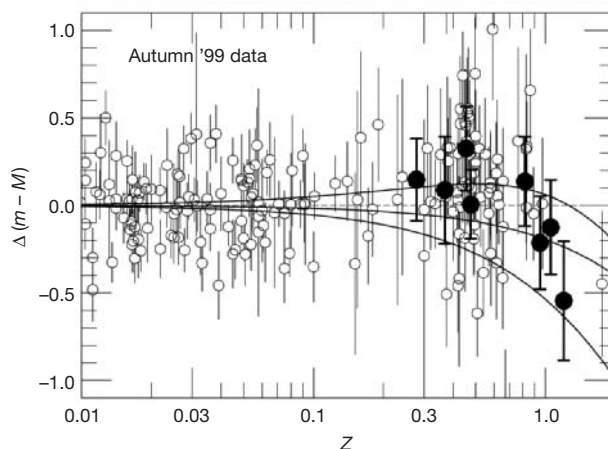


Figure 1 The Hubble diagram (magnitude-redshift diagram) for type Ia supernovae from the High-*z* team²⁹. The vertical axis represents distance modulus ($m-M$), which is a measure of the dimming of the source due to its distance and cosmological expansion. The horizontal axis is redshift, z . The results are plotted in terms of residuals (Δ) obtained by subtracting off the expected behaviour for an empty, undecelerated universe. If this were the correct cosmological model, the results should lie on the straight dashed line. The most recent data (Autumn, 99) are plotted as filled circles with their standard errors; older data are shown as open circles. The diagram also shows the behaviour for three different cosmological models (solid lines): from top to bottom these are the 'concordance model', a model with $\Omega_m = 0.3$ but $\Lambda = 0$, and a model with $\Omega_m = 1$. The data seem to be closest to the concordance model, in which the Universe accelerates at low redshifts but was decelerating at earlier times.

answer. Perhaps there is some reason why the vacuum energy should be exactly zero. Supersymmetric particle theories suggest that this might be possible owing to a cancellation of fermionic and bosonic contributions. However, the cosmological observations suggest this is not correct: there is a non-zero cosmological constant, but it is very small by particle physics standards. In order to get a vacuum energy consistent with the latest data, the cut-off would have to be at the meV scale. Could the right explanation be a supersymmetric theory involving a fundamental energy scale this small?

The realization that our Universe may be accelerating even in its present old age has resulted in a different manifestation of this idea, generically known as quintessence³⁴⁻³⁶. In these models, the cosmological constant (or vacuum energy) is not constant but dynamical, perhaps produced by an evolving scalar field at much lower energies than the GUT scales probably involved in inflation (Box 2). In such models, one can parametrize the behaviour of the field by an effective time-dependent equation of state relating the pressure p and the energy density ρc^2 in the form $p = w(t)\rho c^2$. The case $w = -1$ corresponds to a classical Λ -term, but any value $w < -1/3$ would allow acceleration. Current observations do not allow discrimination between the various possibilities, but that may change in the future. As in the case of primordial inflation, this field may emerge as part of the low-energy effective theory arising from superstring theory, although in this case too the energy scale needs to be explained.

The world after WMAP

The supernova searches were a fitting prelude to the stunning results that emerged from CMB studies that provided definitive evidence for spatial flatness, culminating in the Wilkinson Microwave Anisotropy Probe³⁷ (WMAP) measurements in 2003. The CMB is perhaps the ultimate vehicle for classical cosmology. In looking back to a period

when the Universe was only a few hundred thousand years old, one is looking across most of the observable Universe. This enormous baseline makes it possible to carry out exquisitely accurate surveying. Although the microwave background is very smooth, the COBE satellite did detect small variations in temperature across the sky. These ripples are caused by acoustic oscillations in the primordial plasma. Whereas COBE was only sensitive to long-wavelength waves, WMAP, with its much higher resolution, could probe the higher-frequency content of the primordial roar. The pattern of fluctuations across the sky displays information about characteristic frequencies of the modes of oscillation of the cosmic fireball. The spectrum of the temperature variations displays peaks and troughs³⁸ that contain fantastically detailed data about the basic cosmological parameters described above (and much more); see Fig. 2.

The WMAP data have been greeted with a kind of euphoria by cosmologists as marking the dawn of a new era of ‘precision cosmology’, but perhaps the most striking thing about these results is that they agreed so well with a plethora of previous observations. The process started with COBE³⁹ but there were, for example, strong indications of what WMAP would subsequently confirm in the earlier balloon-based BOOMERANG^{40,41} and MAXIMA^{42–44} experiments, both separately and in combination with each other and with COBE⁴⁵. It is also important to acknowledge the continuing role of smaller-scale interferometric observations, such as CBI⁴⁶ and DASI⁴⁷ (especially in the detection of CMB polarization⁴⁸). These CMB-only constraints were also strengthened by combination with galaxy clustering information^{49,50}. However, because WMAP was able to map the entire sky rather than small patches, the statistical accuracy it achieved over a large range of angular scales brought into the light what had previously been lurking in a thicket of error bars.

Even the preliminary first-year data yield stringent constraints⁵¹ on allowed world models, such as $\Omega_m h^2 = 0.14 \pm 0.02$, $\Omega_b h^2 = 0.024 \pm 0.001$, $h = 0.72 \pm 0.05$ and $\Omega_k = -0.02 \pm 0.02$. These can be strengthened still further by combining the latest WMAP data with supernovae and galaxy clustering measurements⁵². These results are clearly consistent with flat space, and thus fit naturally within the inflationary paradigm, although there is a strange 30%–70% split in the overall energy budget of the Universe, which has yet to be explained by fundamental theory.

The data WMAP has yielded are indeed spectacular, but there are some issues that remain to be resolved. For example, there is a growing realization that the WMAP data contain some strange features^{53–55}, perhaps resulting from some form of foreground contamination⁵⁶. Our own galaxy pollutes the CMB as it is itself a source of dust, synchrotron and free–free emission, some of which appear in microwave frequencies. These must be modelled and corrected before one can see the primordial radiation. This is a difficult task, and it remains to be seen how accurately foreground removal can be accomplished in practice with the relatively simple techniques being used to date.

Whether the residual foreground is sufficiently important to affect the determination of the basic cosmological parameters seems unlikely, but it may disguise more subtle signatures of early Universe physics. One suggestion, for example, has been that the relative lack of large-scale power in the WMAP spectrum may be a consequence of a non-trivial topology for our Universe⁵⁷. This suggestion has been challenged⁵⁸, but it is nevertheless a timely reminder that Einstein’s theory (Box 1) is essentially local and provides no prescription for the large-scale connectivity of space. We could, in principle, be living in a flat universe which is not infinite, but rather finite with a topology like that of a torus.

Future space- and ground-based CMB experiments will focus increasingly not on its temperature variations but on its polarization. The microwave sky is partially polarized, as expected from basic scattering theory. WMAP detected a significant cross-correlation between temperature and polarization fluctuations at a surprisingly high level⁵⁹. This has been interpreted as providing evidence for additional scattering of CMB light owing to the recent reionization of

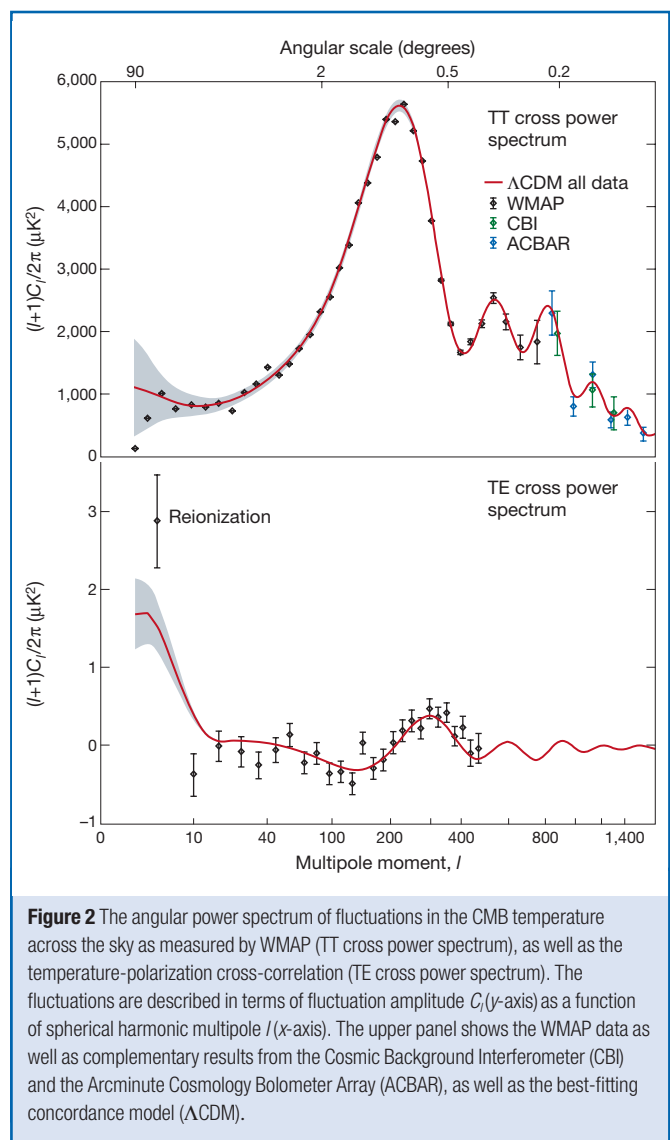


Figure 2 The angular power spectrum of fluctuations in the CMB temperature across the sky as measured by WMAP (TT cross power spectrum), as well as the temperature-polarization cross-correlation (TE cross power spectrum). The fluctuations are described in terms of fluctuation amplitude C_l (y-axis) as a function of spherical harmonic multipole l (x-axis). The upper panel shows the WMAP data as well as complementary results from the Cosmic Background Interferometer (CBI) and the Arcminute Cosmology Bolometer Array (ACBAR), as well as the best-fitting concordance model (Λ CDM).

matter, although it must be stressed it is a preliminary finding. If a relatively weak but polarized source of foreground contamination were in the CMB temperature maps, it might generate a spurious cross-correlation. If clean, detailed maps of the polarization pattern can be made, however, it should prove possible to identify the primordial gravitational waves expected to be produced by inflation⁶⁰.

Structure formation

So far the story has focused on the basic parameters describing an idealized homogeneous and isotropic universe. Of course, our Universe is not really like that. It contains stars, galaxies, clusters and superclusters of galaxies, and a vast amount of messy complexity. Where did this structure come from? The basic idea dates back as far as Jeans¹⁹ and, in a cosmological context, Lifshitz²⁰. A perfectly smooth, self-gravitating fluid with the same density everywhere will stay homogeneous for all time. But any irregularities, however slight, will be amplified by the action of gravity. A small blob with higher than average density will tend to attract material from its surroundings, getting denser still. As it gets denser it attracts yet more matter. The situation is therefore unstable, and the result of this instability is that the blob will turn into a highly concentrated clump held together by its own internal gravitational forces.

Gravitational instability is exponentially fast if the starting configuration is static, but in an expanding universe local gravitational effects must first work to overcome the tendency of the lump to

expand before they can make it contract. The effect of this is that the Universe displays a slower growth of structure, like a power-law function of cosmic time rather than an exponential. The rate of clustering also depends on the overall density of matter, that is, on Ω_m , but not on Λ as the dark energy is assumed not to cluster. The more matter there is, therefore, the faster inhomogeneities grow. Besides inflation, this provided an additional motivation for some cosmologists of the 1980s to lean towards a critical density, that is, $\Omega_m = 1$. Given the tight limits on the allowed baryon abundance, this could only be achieved with the aid of an overwhelmingly predominant distribution of dark non-baryonic material. After pioneering work in numerical simulations of clustering growth, it became clear that the best kind of matter to assist gravitational clustering was cold dark matter (CDM), the required property being that it is slow-moving and easy to collect into structures^{61,62}. The precise nature of this stuff is still unclear, although just about any massive weakly interacting relict of the primordial fireball would do. Supersymmetric theories include fermionic counterparts of the gauge bosons (and the Higgs). If one such particle has a lifetime larger than the age of the Universe, then that could be the dark matter.

For a while, the 'standard' CDM model enjoyed great success, but it is now known to be at odds not just with the basic parameter determinations (which require $\Omega_m < 1$) but also with the statistics of large-scale galaxy clustering. Adoption of the 'concordance' parameter set furnished by WMAP and other measurements gives a ' Λ CDM' model⁶³ that seems to reproduce the large-scale features of the Universe very well (Fig. 3). In this model, the dark energy does not play a major role in the gravitational instability because it does not cluster. The Λ component is important, however, because it causes acceleration that prevents the extremely rapid evolution expected at very late times in models with $\Omega_m = 1$.

There are good theoretical reasons for thinking that galaxies, in statistical terms, roughly trace the mass distribution on large scales up to a bias factor⁶⁴, which can be estimated from the surveys. This is largely borne out by comparison with the latest large redshift surveys, such as the 2dF Galaxy Redshift Survey^{65–68} (2dFGRS; Fig. 4) and the Sloan Digital Sky Survey (SDSS)^{69,70}, whose statistical properties are in good general accord with the gravitational instability picture. The power spectrum of galaxy clustering contains a relict of the acoustic oscillations seen in the CMB angular fluctuations, so comparing galaxy clustering measurements with, for example, the WMAP data, provides a powerful additional consistency check on these models. An independent probe of the bias is also furnished by gravitational lensing measurements, especially using a technique called weak lensing⁷¹, which can provide an independent handle on the underlying mass distribution.

On the scale of individual galaxies, however, things are not so clear. In a nutshell, the idea is that the CDM component begins to cluster in the early Universe, even before recombination (when the baryonic matter is still coupled strongly to radiation). Pools of CDM collect, generating gravitational potential wells into which the gas is subsequently drawn. These pools take the form of extended invisible haloes within which galaxies subsequently form. But the processes involved are complex, and it is hard to predict with certainty what happens inside individual haloes. The gas first heats up, so that it is in gravitational equilibrium with the halo. It then cools through atomic processes and collapses into the centre, fragmenting and forming stars⁶¹. These stars may heat the gas up again, perhaps even expelling some of it from the halo. While this is happening, haloes are continually merging and disrupting each other as gravity generates a hierarchy of structures from the bottom up. While there are also computer codes that can follow the behaviour of baryonic gas alongside the dark matter, there is no available technique for handling star formation, magnetic fields, dust and all the paraphernalia of star formation with similar precision. The best that has been done so far involves so-called semi-analytic models, which use a set of simple *ad hoc* rules to specify when star formation occurs within haloes during this hierarchical

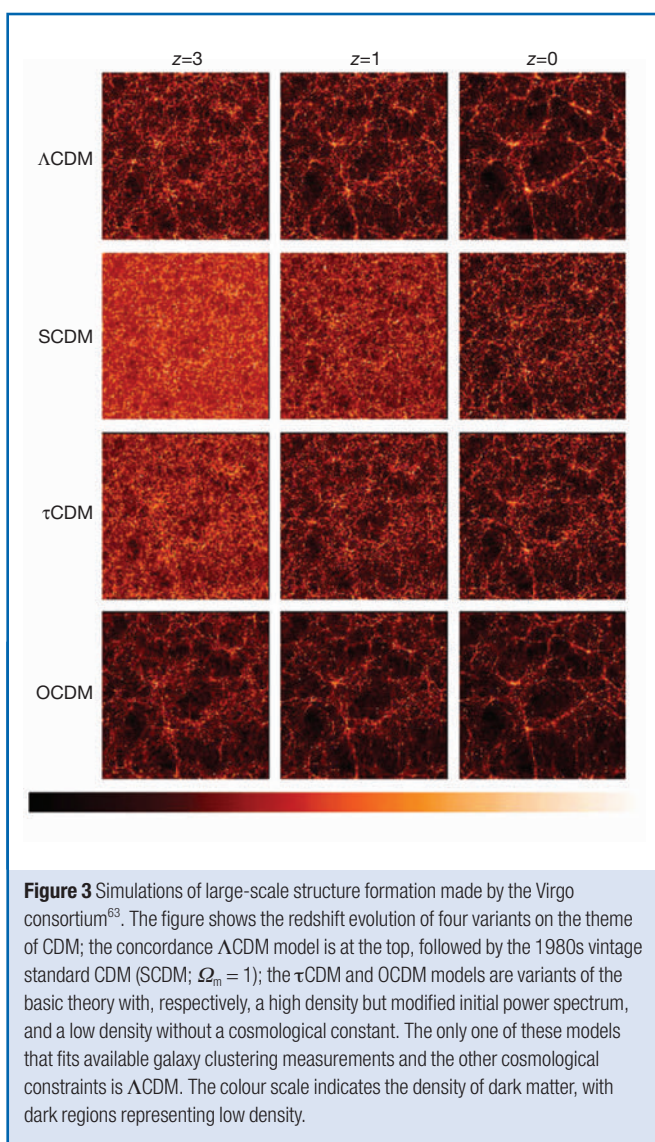


Figure 3 Simulations of large-scale structure formation made by the Virgo consortium⁶³. The figure shows the redshift evolution of four variants on the theme of CDM; the concordance Λ CDM model is at the top, followed by the 1980s vintage standard CDM (SCDM; $\Omega_m = 1$); the τ CDM and OCDM models are variants of the basic theory with, respectively, a high density but modified initial power spectrum, and a low density without a cosmological constant. The only one of these models that fits available galaxy clustering measurements and the other cosmological constraints is Λ CDM. The colour scale indicates the density of dark matter, with dark regions representing low density.

process^{72–75}. This approach has achieved many successes, but also has many shortcomings.

One of the basic problems concerns the distribution of dark matter in galaxies. High-resolution numerical experiments⁷⁶ suggest that the density profile $\rho(r)$ of dark matter haloes, as a function of distance r from the halo centre, is expected to follow the universal form:

$$\rho(r) = Ar^{-\gamma} \left(1 + \frac{r}{r_s}\right)^{-3} \quad (2)$$

where A is a normalization parameter and r_s is a characteristic scale related to the mass and degree of concentration of the halo. The slope γ is measured in simulations to be $\gamma \approx 1.2 \pm 0.3$. These experiments involve only gravity and cold collisionless matter, but even these results are somewhat controversial^{77–82}.

Observations seem to be in disagreement with the predictions of equation (2), although this too is controversial^{83–85}. The concentration of CDM seems to be low in galaxies with a low surface brightness of stellar emission. The steep 'cusp' in the centre is not usually seen in such galaxies: the core seems to have a flatter slope. There also seems to be a puzzling dearth of dark matter in some elliptical galaxies⁸⁶. Moreover, the dark matter in galaxy haloes is expected to be highly clumped, leading to large numbers of dwarf haloes. If these objects formed stars as efficiently as 'normal' galaxies, then they would exceed the observed number of dwarf galaxies by a large factor. The CDM clumpiness is also expected to cause baryons to lose angular

The undiscovered cosmos

So this is the current state of the Universe, or at least of our understanding of parts of it. We seem to have established a 'standard' cosmological model dominated by dark energy and dark matter, with a tiny flavouring of the baryonic matter from which stars, planets and ourselves are made. This standard model accounts for many precise observations and has been hailed as a spectacular triumph. And so it is. But this progress should not distract us from the fact that modern cosmology also has a number of serious shortcomings that may take a long time to remedy. These difficulties can be summarized in a series of interrelated questions, most of them very basic, to which we still do not have answers, but where a number of upcoming observational programmes⁹⁶ may yield the crucial clues.

Is general relativity right? In a sense we already know that the answer to this question is 'no'. At sufficiently high energies (near the Planck scale), we expect classical relativity to be replaced by a quantum theory of gravity. But it is not just in the early Universe that departures from general relativity are possible.

The different estimates of $\Omega_k \approx 0$ (for example, from WMAP), $\Omega_\Lambda \approx 0.7$ (from supernovae) and $\Omega_m \approx 0.3$ (from galaxy clustering) do seem to satisfy equation (1). This, together with the agreement between stellar ages and the dynamical age of the Big Bang, shows that we at least have a consistent picture based on general relativity. Whether this consistency will survive the test of time remains to be seen.

Is the Universe finite or infinite? If we accept the Big Bang theory then our observable Universe began a finite time in the past. The finite speed of light means that the part of it we can expect to see must be finite. But what happens beyond our horizon is not known, even if our observable patch is well described by the concordance model. With a cosmological constant, our accelerating Universe might continue speeding up for the indefinite future, but if the dark energy were dynamical the far future depends on its eventual ground state, which is unknown. Our Universe may not be infinite in duration, even if it is currently accelerating.

What is the dark energy? The question here is twofold. One part is whether the dark energy is of the form of an evolving scalar field, such as quintessence or its variants, or whether it really is constant, as in Einstein's original version. This may well be answered by future observational studies, such as the SNAP satellite (currently awaiting budget approval in the USA), which will observe not dozens but thousands of supernovae, and may be able to detect changes in the parameter w with redshift. The second part is whether dark energy can be understood in terms of fundamental theory. This is a more open question. At the least, it would require an understanding of the field(s) involved from a particle physics viewpoint. But it would also require an understanding of why the expected divergence of one-loop fluctuations is suppressed. I think it is safe to say that we are still far from such an understanding. Perhaps string theory holds the key.

What is the dark matter? Around 30% of the mass in the Universe is thought to be in the form of dark matter, but we do not know what it is — other than it is meant to be cold. As far as gravity is concerned, one cold particle is much the same as another. There is no prospect for learning about the nature of the CDM particles through astronomical means, unless they decay into radiation or some other identifiable particles. Experimental attempts to detect the dark matter directly are pushing back the limits of technology, but it would have to be a long shot for them to succeed in such a large parameter space.

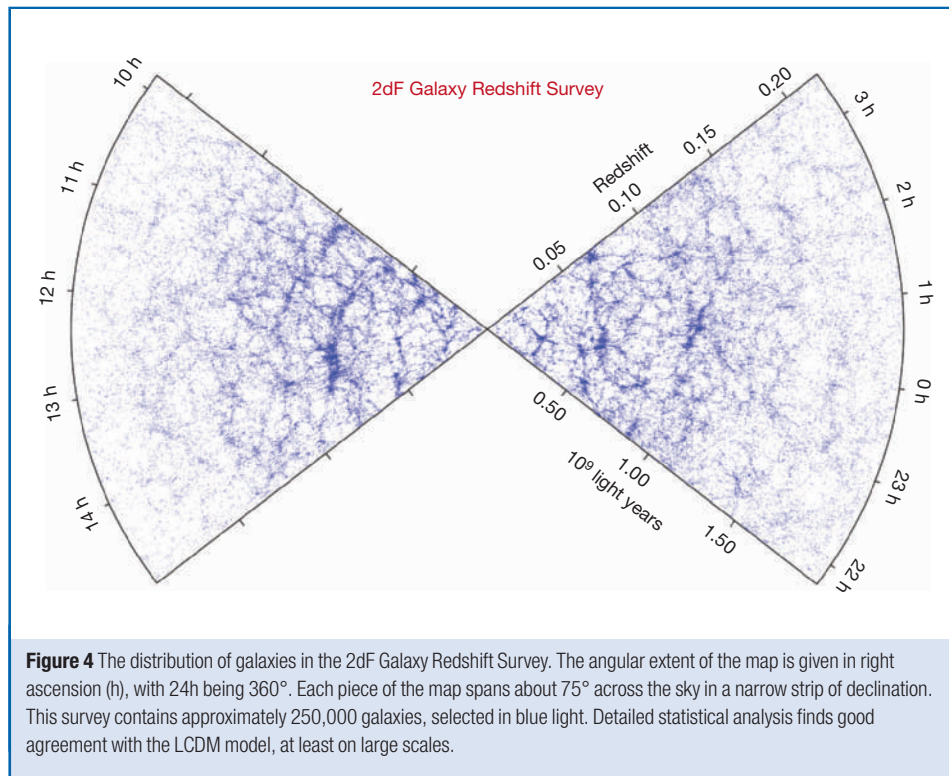


Figure 4 The distribution of galaxies in the 2dF Galaxy Redshift Survey. The angular extent of the map is given in right ascension (h), with 24h being 360°. Each piece of the map spans about 75° across the sky in a narrow strip of declination. This survey contains approximately 250,000 galaxies, selected in blue light. Detailed statistical analysis finds good agreement with the LCDM model, at least on large scales.

momentum by dynamical friction as they collapse. Galaxy disks are then predicted to be about a factor of five smaller than observed. All these conclusions may be affected to a greater or lesser extent by problems with numerical resolution.

Overall, the semi-analytic models struggle to explain the observed luminosity function of galaxies⁸⁷: gas cooling (and consequent star formation) seems to be so efficient that the models predict too many very luminous galaxies. Some process, such as feedback from active galactic nuclei, supernovae, or stellar winds, may regulate the supply of cold gas by providing feedback that heats up or ejects the material falling into a galaxy, but the details of how this happens remain obscure. More dramatically, the role of a central supermassive black hole in the evolution of galaxies is yet to be fully elucidated. Perhaps accretion onto such objects causes violent outflows that expel at least some of the cold gas. The problem of galaxy formation is so complex that it is unlikely to be solved from first principles. What is more likely is that observations will pin down the range of possibilities to enable better models to be constructed.

A key question on the route to a fuller picture is how the whole process of galaxy formation got under way. After the production of the CMB radiation, baryonic matter is thought to have recombined into a neutral gas which cooled as the Universe expanded. At some point, something happened to inject sufficient energy into this gas to heat it up again to the point of ionization. Circumstantial evidence for when this happened is furnished by the WMAP polarization measurements⁵⁹, as well as by more direct measurements of the intergalactic medium through lines-of-sight to distant quasars^{88,89}. We do not know what kind of object — primordial stars, active galaxies or quasars — produces the first flash of ionizing radiation^{90,91}, and we do not know the extent to which the reionization occurs smoothly or in lumps. And how does the epoch of reionization of the Universe relate to the onset of star formation in galaxies^{92,93}? Answers to these questions will require observing facilities across a whole range of wavelengths, because light produced in the early Universe is seen at much longer wavelengths by present-day observers. Moreover, indications are that as much as half the stellar light⁹⁴ produced as galaxies form and evolve has been obscured by dust, to be re-radiated in the infrared and sub-millimetre regions of the spectrum⁹⁵.

Did inflation really happen? The success of concordance cosmology is largely founded on the appearance of ‘Doppler peaks’ in the CMB fluctuation spectrum. These arise from acoustic oscillations in the primordial plasma that have particular statistical properties consistent with them having been generated by quantum fluctuations in the scalar field driving inflation. This is circumstantial evidence in favour of inflation happening, but it is perhaps not strong enough to give a definitive answer. The definitive proof of inflation is probably the existence of a stochastic gravitational wave background; in the ekpyrotic universe, for example, the predicted background is very small. The identification and extraction of this may be possible using future polarization-sensitive CMB studies, such as the Planck Surveyor, even before direct experimental gravitational wave detectors of sufficient sensitivity, such as LISA, become available.

How and when did galaxy formation happen? The field of galaxy formation has changed immeasurably over the past decade. Interplay between theory, modelling and observation is now so intense that the boundaries have become blurred. The imminent arrival of the Atacama Large Millimetre Array (ALMA) for submillimetre observations of dust emission at high redshift and, in the longer term, 100-m-class optical telescopes to image very distant star-forming regions, and the Square Kilometre Array (SKA) for sensitive radio measurements, are just some of the developments on the horizon. SKA in particular will enable observations to reach back past the epoch of galaxy formation and into the cosmic dark ages before there were any stars, and also probe directly the supply of cold gas through its 21-cm emission. Provided that the necessary investment in instrumentation is made, this is one question where a firm, empirically based answer is likely. □

doi:10.1038/nature03282

1. Einstein, A. The gravitational field equations [in German]. *Sitz. Ber. Preuss. Akad. Wiss.* **844**–847 (1915).
2. Friedmann, A. On the curvature of space [in German]. *Z. Phys.* **10**, 377–386 (1922).
3. Lemaitre, G. A homogenous universe of constant mass and increasing radius accounting for the radial velocity of the extragalactic nebulae. *Ann. Soc. Sci. Brux. A* **47**, 49–59 (1929); [in English] *Mon. Not. R. Astron. Soc.* **91**, 483–490 (1931).
4. Hubble, E. A relation between distance and radial velocity among extragalactic nebulae. *Proc. Natl Acad. Sci.* **15**, 168–173 (1929).
5. Penzias, A. A. & Wilson, R. W. Measurement of excess antenna temperature at 4080 MHz. *Astrophys. J.* **142**, 419–421 (1965).
6. Dicke, R. H., Peebles, P. J. E., Roll, P. G. & Wilkinson, D. T. Cosmic blackbody radiation. *Astrophys. J.* **142**, 414–419 (1965).
7. Mather, J. C. *et al.* Measurement of the cosmic background spectrum by the COBE FIRAS instrument. *Astrophys. J.* **420**, 439–444 (1994).
8. Gamow, G. Expanding universe and the origin of elements. *Phys. Rev.* **70**, 572–573 (1946).
9. Alpher, R. A. & Herman, R. C. Evolution of the universe. *Phys. Rev.* **73**, 803–804 (1948).
10. Guth, A. H. Inflationary Universe: A possible solution to the horizon and flatness problems. *Phys. Rev. D* **23**, 347–356 (1981).
11. Albrecht, A. & Steinhardt, P. J. Cosmology for grand unified theories with radiatively induced symmetry breaking. *Phys. Rev. Lett.* **48**, 1220–1223 (1982).
12. Linde, A. D. A new inflationary universe scenario: A possible solution of the horizon, flatness, homogeneity, isotropy and primordial monopole problems. *Phys. Lett. B* **108**, 389–393 (1982).
13. Linde, A. D. Chaotic inflation. *Phys. Lett. B* **129**, 177–181 (1983).
14. Kaluza, T. On the unification problem of physics [in German]. *Sitz. Ber. Preuss. Akad. Wiss.* 966–972 (1921).
15. Klein, O. Quantum theory and five-dimensional relativity theory [in German]. *Z. Phys.* **37**, 895–906 (1926).
16. Khoury, J., Ovrut, B. A., Steinhardt, P. J. & Turok, N. Ekpyrotic universe: Colliding branes and the origin of the hot big bang. *Phys. Rev. D* **64**, 13522 (2001).
17. Khoury, J., Ovrut, B. A., Steinhardt, P. J. & Turok, N. From big crunch to big bang. *Phys. Rev. D* **65**, 086007 (2002).
18. Randall, L. & Sundrum, R. An alternative to compactification. *Phys. Rev. Lett.* **83**, 4690–4693 (1999).
19. Jeans, J. H. The stability of spiral nebulae. *Phil. Trans. R. Soc. Lond. A* **199**, 1–53 (1902).
20. Lifshitz, E. M. On the gravitational instability of the expanding universe. *Sov. Phys. JETP* **10**, 116–122 (1946).
21. Linde, A. D. Scalar field fluctuations in the expanding universe and the new inflationary universe scenario. *Phys. Lett. B* **116**, 335–339 (1982).
22. Guth, A. H. & Pi, S.-Y. Fluctuations in the new inflationary universe. *Phys. Rev. Lett.* **49**, 1110–1113 (1982).
23. Starobinsky, A. A. Spectrum of relic gravitational radiation and the early state of the universe. *Sov. Phys. JETP Lett.* **30**, 682–685 (1979).
24. Coles, P. & Ellis, G. F. R. The case for an open universe. *Nature* **370**, 609–615 (1994).
25. Efstathiou, G., Sutherland, W. J. & Maddox, S. J. The cosmological constant and cold dark matter. *Nature* **348**, 705–706 (1990).
26. Perlmutter, S. *et al.* Measurements of omega and lambda from 42 high-redshift supernovae. *Astrophys. J.* **517**, 565–586 (1999).
27. Riess, A. G. *et al.* Observational evidence from supernovae for an accelerating universe and a cosmological constant. *Astron. J.* **116**, 1009–1038 (1998).
28. Riess, A. G. *et al.* The farthest known supernova: Support for an accelerating universe and a glimpse of the epoch of deceleration. *Astrophys. J.* **560**, 49–71 (2001).
29. Tonry, J. L. *et al.* Cosmological results from high-*z* supernovae. *Astrophys. J.* **594**, 1–24 (2003).
30. Freedman, W. L. *et al.* Final results from the Hubble Space Telescope key project to measure the Hubble constant. *Astrophys. J.* **553**, 47–72 (2001).
31. Carretta, E., Gratton, R. G., Clementini, G. & Flaviaio, F. P. Distances, ages and epoch of formation of globular clusters. *Astrophys. J.* **533**, 215–235 (2000).
32. Casimir, H. B. G. On the attraction between two perfectly conducting plates. *Proc. K. Ned. Akad. Wet.* **51**, 793–795 (1948).
33. Weinberg, S. The cosmological constant problem. *Rev. Mod. Phys.* **6**, 1–22 (1989).
34. Wang, L., Caldwell, R. R., Ostriker, J. P. & Steinhard, P. J. Cosmic concordance and quintessence. *Astrophys. J.* **530**, 17–35 (2000).
35. Sahni, V. The cosmological constant problem and quintessence. *Class. Quant. Grav.* **19**, 3435–3448 (2002).
36. Peebles, P. J. E. & Ratra, B. The cosmological constant and dark energy. *Rev. Mod. Phys.* **75**, 559–606 (2003).
37. Bennett, C. L. *et al.* First-year Wilkinson Microwave Anisotropy Probe (WMAP) observations. *Astrophys. J. Suppl.* **148**, 1–27 (2003).
38. Hinshaw, G. *et al.* First-year Wilkinson Microwave Anisotropy Probe (WMAP) observations: The angular power spectrum. *Astrophys. J. Suppl.* **148**, 135–159 (2003).
39. Smoot, G. F. *et al.* Structure in the COBE differential microwave radiometer first year maps. *Astrophys. J.* **396**, L1–L4 (1992).
40. de Bernardis, P. *et al.* A flat universe from high-resolution maps of the cosmic microwave background radiation. *Nature* **420**, 763–765 (2000).
41. Lange, A. E. *et al.* Cosmological parameters from the first results of BOOMERANG. *Phys. Rev. D* **63**, 042001 (2001).
42. Hanany, S. *et al.* MAXIMA-1: A measurement of the cosmic microwave background anisotropy on angular scales 10 arc minutes to 5 degrees. *Astrophys. J.* **545**, L5–L9 (2000).
43. Balbi, A. *et al.* Constraints on cosmological parameters from MAXIMA-1. *Astrophys. J.* **545**, L1–L4 (2000).
44. Lee, A. *et al.* A high spatial resolution analysis of the MAXIMA-1 cosmic microwave background anisotropy data. *Astrophys. J.* **561**, L1–L4 (2001).
45. Jaffe, A. H. Cosmology from MAXIMA-1, BOOMERANG, and COBE-DMR cosmic microwave background observations. *Phys. Rev. Lett.* **86**, 3475–3479 (2001).
46. Padin, S. *et al.* First intrinsic anisotropy measurements with the cosmic background imager. *Astrophys. J.* **549**, L1–L5 (2001).
47. Halverson, N. W. *et al.* Degree Angular Scale Interferometer first results: A measurement of the cosmic microwave background angular power spectrum. *Astrophys. J.* **568**, 38–45 (2002).
48. Kovac, J. M. *et al.* Detection of polarization in the cosmic microwave background using DASI. *Nature* **468**, 46–51 (2002).
49. Efstathiou, G. Evidence for a non-zero and a low matter density from a combined analysis of the 2dF Galaxy Redshift Survey and cosmic microwave background anisotropies. *Mon. Not. R. Astron. Soc.* **330**, L29–L35 (2002).
50. Lahav, O. *et al.* The 2dF Galaxy Redshift Survey: The amplitudes of fluctuations in the 2dFGRS and the CMB, and the implications for galaxy biasing. *Mon. Not. R. Astron. Soc.* **333**, 961–968 (2002).
51. Spergel, D. N. *et al.* First-year Wilkinson Microwave Anisotropy Probe (WMAP) observations: Determination of cosmological parameters. *Astrophys. J. Suppl.* **148**, 175–194 (2003).
52. Bridle, S. L., Lahav, O., Ostriker, J. P. & Steinhardt, P. J. Precision cosmology? Not just yet. *Science* **299**, 1532–1536 (2003).
53. Eriksen, H. K., Hansen, F. K., Banday, A. J., Gorski, K. M. & Lilje, P. B. Asymmetries in the cosmic microwave background anisotropy field. *Astrophys. J.* **605**, 14–20 (2004).
54. Vielva, P., Martinez-Gonzalez, E., Barreiro, R. B., Sanz, J. L. & Cayon, L. Detection of non-gaussianity in the Wilkinson Microwave Anisotropy Probe first-year data using spherical wavelets. *Astrophys. J.* **609**, 22–34 (2004).
55. Coles, P., Dineen, P. J., Earl, J. & Wright, D. Phase correlations in cosmic microwave background temperature maps. *Mon. Not. R. Astron. Soc.* **350**, 989–1004 (2004).
56. De Oliveira-Costa, A. *et al.* The quest for microwave foreground X. *Astrophys. J.* **606**, L89–L92 (2004).
57. Luminet, J. P., Weeks, J. R., Riazuelo, A., Lehoucq, R. & Uzan, J.-P. Dodecahedral space topology as an explanation for the weak wide-angle temperature correlations in the cosmic microwave background. *Nature* **425**, 593–595 (2003).
58. Cornish, N. J., Spergel, D. N., Starkman, G. D. & Komatsu, E. Constraining the topology of the universe. *Phys. Rev. Lett.* **92**, 201302 (2004).
59. Kogut, A. *et al.* First-year Wilkinson Microwave Anisotropy Probe (WMAP) observations: Temperature-polarization correlation. *Astrophys. J. Suppl.* **148**, 161–173 (2003).
60. Kosowsky, A. Introduction to microwave background polarization. *New Astron. Rev.* **48**, 157–168 (1999).
61. Blumenthal, G. R., Faber, S. M., Primack, J. R. & Rees, M. J. Formation of galaxies and large-scale structure with cold dark matter. *Nature* **311**, 517–525 (1984).
62. Davis, M., Efstathiou, G., Frenk, C. S. & White, S. D. M. The evolution of large-scale structure in a universe dominated by cold dark matter. *Astrophys. J.* **292**, 371–394 (1985).
63. Jenkins, A. *et al.* Evolution of structure in CDM universes. *Astrophys. J.* **499**, 20–40 (1998).
64. Coles, P. Galaxy formation with a local bias. *Mon. Not. R. Astron. Soc.* **262**, 1065–1075 (1993).
65. Colless, M. *et al.* The 2dF Galaxy Redshift Survey: Spectra and redshifts. *Mon. Not. R. Astron. Soc.* **328**, 1039–1063 (2001).
66. Percival, W. J. *et al.* The 2dF Galaxy Redshift Survey: The power spectrum and the matter content of the universe. *Mon. Not. R. Astron. Soc.* **327**, 1297–1306 (2001).
67. Peacock, J. A. *et al.* A measurement of the cosmological mass density from clustering in the 2dF Galaxy Redshift Survey. *Nature* **410**, 169–163 (2001).
68. Norberg, P. D. *et al.* The 2dF Galaxy Redshift Survey: Luminosity dependence of galaxy clustering. *Mon. Not. R. Astron. Soc.* **328**, 64–70 (2001).
69. Abazajian, K. *et al.* The first data release of the Sloan Digital Sky Survey. *Astron. J.* **126**, 2081–2086 (2003).
70. Zehavi, I. *et al.* Galaxy clustering in early Sloan Digital Sky Survey redshift data. *Astrophys. J.* **571**, 172–190 (2002).

71. Mellier, Y. Probing the Universe with weak lensing. *Annu. Rev. Astron. Astrophys.* **37**, 127–189 (1999).
72. Kauffmann, G., White, S. D. M. & Guiderdoni, B. The formation and evolution of galaxies within merging dark matter haloes. *Mon. Not. R. Astron. Soc.* **264**, 201–218 (1993).
73. Somerville, R. S. & Primack, J. R. Semi-analytic modelling of galaxy formation: the local universe. *Mon. Not. R. Astron. Soc.* **264**, 201–218 (1993).
74. Baugh, C. M., Cole, S., Frenk, C. S. & Lacey, C. G. The epoch of galaxy formation. *Astrophys. J.* **498**, 504–521 (1998).
75. Cole, S., Lacey, C. G., Baugh, C. M. & Frenk, C. S. Hierarchical galaxy formation. *Mon. Not. R. Astron. Soc.* **319**, 168–204 (2000).
76. Navarro, J. F., Frenk, C. S. & White, S. D. M. A universal density profile from hierarchical clustering. *Astrophys. J.* **490**, 493–508 (1997).
77. Moore, B., Quinn, T., Governato, F., Stadel, J. & Lake, G. Cold collapse and the core catastrophe. *Mon. Not. R. Astron. Soc.* **310**, 1147–1152 (1999).
78. Ghigna, S. *et al.* Density profiles and substructure of dark matter halos: Converging results and ultra-high numerical resolution. *Astrophys. J.* **544**, 616–628 (2000).
79. Jing, Y.-P. & Suto, Y. The density profiles of the dark matter halo are not universal. *Astrophys. J.* **529**, L69–L72 (2000).
80. Klypin, A., Kravtsov, A. V., Bullock, J. S. & Primack, J. R. Resolving the structure of cold dark matter halos. *Astrophys. J.* **554**, 903–915 (2001).
81. Fukushige, T., Kawai, A. & Makino, J. Structure of dark matter halos from hierarchical clustering III. Shallowing of the inner cusp. *Astrophys. J.* **606**, 625–634 (2004).
82. Navarro, J. F. *et al.* The inner structure of CDM haloes — III. Universality and asymptotic slopes. *Mon. Not. R. Astron. Soc.* **349**, 1039–1051 (2004).
83. van den Bosch, F. C., Robertson, B. E., Dalcanton, J. J. & de Blok, W. J. G. Constraints on the structure of dark matter halos from the rotation curves of low surface brightness galaxies. *Astron. J.* **119**, 1579–1591 (2000).
84. de Blok, W. J. G., McGaugh, S. S., Bosma, A. & Rubin, V. C. Mass density profiles of low surface brightness galaxies. *Astron. J.* **122**, 2396–2347 (2001).
85. van den Bosch, F. C. & Swaters, R. A. Dwarf galaxy rotation curves and the core problem of dark matter haloes. *Mon. Not. R. Astron. Soc.* **325**, 1017–1038 (2001).
86. Romanowsky, A. *et al.* A dearth of dark matter in ordinary elliptical galaxies. *Science* **301**, 1696–1698 (2003).
87. Benson, A. J. *et al.* What shapes the luminosity function of galaxies? *Astrophys. J.* **599**, 38–49 (2003).
88. Becker, R. H. *et al.* Evidence for reionization at $z \approx 6$: Detection of a Gunn–Peterson trough in a $z=6.28$ quasar. *Astron. J.* **122**, 2850–2857 (2001).
89. Fan, X. *et al.* Evolution of the ionizing background and the epoch of reionization from the spectra of $z=6$ quasars. *Astron. J.* **123**, 1247–1257 (2002).
90. Gnedin, N. Y. Cosmological reionization by stellar sources. *Astrophys. J.* **535**, 530–554 (2000).
91. Barkana, R. & Loeb, A. In the beginning: The first sources of light and the reionization of the universe. *Phys. Rep.* **349**, 125–238 (2001).
92. Abel, T., Bryan, G. & Norman, M. L. The formation of the first star in the universe. *Science* **295**, 93–98 (2002).
93. Bromm, V. & Larson, R. B. The first stars. *Annu. Rev. Astron. Astrophys.* **42**, 79–118 (2004).
94. Madau, P., Pozzetti, L. & Dickinson, M. The star formation histories of field galaxies. *Astrophys. J.* **498**, 106–116 (1998).
95. Blain, A. W., Smail, I., Ivison, R. J. & Kneib, J.-P. The history of star formation in dusty galaxies. *Mon. Not. R. Astron. Soc.* **302**, 632–648 (1999).
96. Coles, P. The future of extragalactic observations. *Class. Quant. Grav.* **19**, 3539–3549 (2002).
97. Einstein, A. Cosmological considerations of the general theory of relativity [in German]. *Sitz. Ber. Preuss. Akad. Wiss.* 142–152 (1917); English translation in *The Principle of Relativity* (eds Lorentz, H. A., Einstein, A., Minkowski, H. & Weyl, H.) 177–188 (Methuen, London, 1950).

Competing interests statement The author declares that he has no competing financial interests.

A theory of everything?

In his later years, Einstein sought a unified theory that would extend general relativity and provide an alternative to quantum theory. There is now talk of a 'theory of everything' (although Einstein himself never used the phrase). Fifty years after his death, how close are we to such a theory?

have learned from the history of physics that the most spectacular developments usually come when the existing theoretical insights create conflict, when observed phenomena seem to defy logic. This happened when Max Planck investigated a problem in the statistical features of radiation — which led to the foundation of quantum mechanics. It also happened when Einstein asked questions about observers moving at the speed of light — which led to the theory of relativity. On a more modest scale, a problem with neutral currents in the weak interactions of particles led to the prediction of the charm quark.

History has a habit of repeating itself, but not in a predictable manner. So, even if we are facing similar problems today, where will they lead us? Now, our most prominent difficulty is the reconciliation of the general theory of relativity with quantum mechanics. And one of the real paradoxes is the tiny value of the cosmological constant. There are also conceptual difficulties with black holes, or, more precisely, with what happens when the gravitational force exhibits its notorious instability.

I once planned to write a paper entitled '200 incorrect theories to explain the cosmological constant', with 200 references, but I never had the patience to read all those papers. I am convinced that this problem, and that of the black holes, will require a much more revolutionary approach. Many of my colleagues, in particular string theorists, expect that the ultimate laws of physics will contain a kind of logic that is even more mysterious and alien than that of quantum mechanics. I, however, will only be content if the logic is found to be completely straightforward. I suspect that the favoured interpretation of quantum mechanics will have to be revised. I am not saying that quantum mechanics is wrong or incomplete. But I do think that an ultimate theory will not have any stochastic elements: I side with Albert Einstein, who always suspected that nature's true equations would not allow for gambling. **Gerard 't Hooft**

Physics as we know it is undergoing a profound transformation. The focus of twentieth-century physics — the structure of matter — is shifting to the structure of the Universe. Physicists are asking (and hoping to answer) new kinds of questions. Instead of asking what the laws or equations of physics are and how we solve them, we ponder why the laws of physics are as they are, and ask whether they could be otherwise. Could things be different in other parts of a much bigger universe than any we have ever imagined? Why are the laws and constants of nature so well-tuned to the existence of life? Is it accident or is there a pattern?

There are contrasting views on these questions. Theoretical physicists have always hoped that the underlying laws of physics — the laws of particle physics — would be uniquely determined by the internal consistency of some particularly simple mathematical theory. Not only would the theory explain why the proton and neutron are about 1,800 times heavier than the electron, but the theory would also explain itself: no other theory is possible. Moreover, according to this view the fantastic coincidences that seem

incredibly fine tuned, just so that life can exist, are exactly that — just lucky coincidences.

Many physicists hoped that string theory would be the mathematical 'silver bullet' that would uniquely explain our world. But the more we learn about cosmology and the more we learn about string theory, the less likely this seems. Experimental cosmology has given us reasons to believe two things: that the Universe is vastly bigger than the 10 billion light-years that we can see with our astronomical instruments; and that at least one constant of nature, the so-called cosmological constant, is absurdly fine tuned (to more than one hundred decimal places) to be in the range where galaxies, stars and life can form. And, in a reversal of fortune that stunned string theorists, it seems that their own theory has so many solutions that there is an incredible 'landscape' of possibilities.

The competing view — that physicists generally hate, but might have to take seriously — is that the Universe is tremendously big and filled with smaller 'pocket universes', each with its own peculiar elementary particles, forces and constants of nature. This seems to be both what observational cosmology and string theory are pushing us towards. If so, the explanation of some aspects of nature would then be that life as we know it can only exist in those regions where the conditions are suitable. Who is right? Hopefully time will tell.

Leonard Susskind

Einstein famously devoted the second half of his career to attempting to unify the forces of nature. To Einstein, this meant unifying electromagnetism and gravitation, the only forces he regarded as truly fundamental. Nowadays, we include weak interactions and the strong force as coequal partners. We aim to unify general relativity — Einstein's theory of gravity — with quantum field theory, the framework in which we understand the other three forces.

Does such a unified theory exist? Can we find it? Or could the search go on forever — with each step forward raising a new riddle? If there is a unified field theory and we find it, can we do enough decisive experiments and calculations to confirm that it is right? Can a unified field theory be used to compute the dimensionless constants that we observe in nature, or do the values of these constants depend on the choice of a solution of the unified field theory? Are they different in different parts of the Universe?

One of the few things we do know is that, with string theory, theoretical physicists have stumbled upon a theory that looks like it might be the unified field theory. It modifies the notions of quantum field theory and forces upon us the unification of gravity and quantum theory. It can naturally encompass the standard model of particle physics. And it always seems to defy full understanding, no matter how much progress is made. In thirty-something years of intense work, string theory has generated an amazing panoply of rich ideas, with wide influence in different areas of physics and mathematics. But it continues to baffle its practitioners.

Edward Witten

Cosmologists have already reached a reasonable understanding of the global evolution of the Universe and the formation of cosmic structure — notwithstanding three fundamental yet hypothetical components of the theory, namely dark matter, the cosmological constant and inflation (or the scalar field that induces inflation). How then might we be concerned with a ‘theory of everything’?

We probably do not need such a theory to answer the dark matter problem — we hope that experiments will reveal the nature of this matter directly. The existence of the cosmological constant, or vacuum energy, has fundamental implications for physics. If a theory of everything were to tell us nothing specific about this energy, that would perhaps be the most important result. We would then not need to worry about its arbitrary value (and could appeal to the anthropic principle without hesitation). Or perhaps the theory of everything contains a compromise solution, with a dynamically varying vacuum energy.

Inflation provides the seed of cosmic structure and fits recent data beautifully. But the model is still poorly understood, and this problem is most closely connected with the theory of everything. But progress here should be driven by developments in particle physics, towards a model that can predict the correct magnitude of cosmological fluctuations. Cosmology is, unfortunately, not useful in constructing a theory of everything — although a theory of everything would solve some important problems in cosmology. We watch and wait!

Masataka Fukugita

The quest for a unifying theory is an ambitious task. The ultimate goal is a simple, elegant, unifying theory — one that can be used to predict the result of any conceivable experiment. But nature manifests only a fraction of the simplicity that such a theory is supposed to contain. A unified theory would have to connect a spare formulation to our apparently far more complicated world.

This means that even if simple principles underlie physical reality, more elaborate theoretical ideas are needed to relate these principles to actual physical phenomena. That is not necessarily a bad thing: one of the most interesting challenges in physics is to figure out such relationships. But it does mean that a truly unified theory is likely to contain some complicated elements. For example, symmetry is essential to the simplicity of the beautiful theoretical description of the weak nuclear force. But that symmetry is broken by the state of the Universe in which we live. The underlying theory is simple and unifies the weak force and electromagnetism, but the reality is more complicated. We need new experimental results to complete the story.

String theory, which accommodates both quantum mechanics and gravity, aims to be a unified theory. But to derive the properties of the world in which we live, theorists need to make assumptions about various parameters. Originally, string theorists hoped string theory would dictate these parameters. But this is looking increasingly unlikely. It is far more probable that there aren’t enough rules to predict all the parameters of our world.

Physics has changed dramatically over the years as we have learned about physical principles that apply at higher energies and smaller length scales. New information from upcoming experiments, such as those at the Large Hadron Collider at CERN in Switzerland, should help us discover further guiding principles. Progress will depend on relating intriguing features of the results of such experiments to the consequences of various theories, such as string theory. I don’t know if we’ll ever know all the answers or find a single predictive unifying theory. But I’m confident that we will continue to make progress towards a deeper understanding of nature’s fundamental laws.

Lisa Randall

I dislike the phrase ‘a theory of everything’. It is arrogant and suggests that all that remains to do is to tie the world, as we know it, up in a single theory. But I do believe that progress will be made in the discovery of a deeper and more unified understanding of the phenomena so far discovered.

Can quantum theory and general relativity be unified? The first step towards this goal is to invent a common language and theory in which the principles of general relativity and quantum theory work together consistently. The good news is that a theory of this description has been found: it is called loop quantum gravity. This is not to say that all questions about this theory have been answered, or that we know that it describes nature. But it is a consistent, well-defined theory that does follow from combining the principles of both quantum theory and general relativity. In particular, it realizes perfectly the basic principle embodied in Einstein’s theory that space and time are dynamical, and evolve with matter, rather than being a fixed background. It makes a major new prediction which is that space, like atoms, has a discrete structure and it predicts the details of that structure. This discreteness might be tested in real experiments, which are now underway.

A further step would be to discover a common origin for the geometry of space-time and quantum phenomena, so as to truly unify them. In such a theory there would be at first nothing recognizable as either space or quantum physics: these would emerge as approximate descriptions, just as temperature and pressure emerge from the statistics of the motions of large numbers of atoms. Such a theory does not exist but some first steps have been taken towards formulating it.

Lee Smolin

Two years after Einstein completed the special theory of relativity, he was already at work on an extension of this theory which included gravitation. He quickly seized on the equivalence of gravitational and inertial mass as the key to understanding gravitation. To take account of this ‘equivalence principle’, the special theory of relativity had to give way to a generalized theory of relativity and gravitation, in which inertia and gravity are united and no privileged frames of reference exist.

By 1915, Einstein had a general theory in which all space-time structures become dynamical fields. This is quite a remarkable conclusion. All other successful quantum theories — in particular, non-relativistic quantum mechanics and special relativistic quantum field theory — have incorporated some kinematical background space-time structure, a stage on which the dramas of dynamics are enacted. Now, there is no kinematics independent of dynamics: in this sense, general relativity is a background-independent theory.

Attempts to create a quantum theory of gravity are thus faced with the following problem: must we give up background independence to quantize gravity? This has been the choice made (so far, at least) by perturbative string theorists, who introduce a background — flat space-time of ten, eleven or however many dimensions — and then apply a variant of the procedures used in (special relativistic) quantum field theory to quantize the string in this background (the excited states of the quantized string represent particles).

But loop quantum gravity theorists maintain that it is not only possible, but mandatory to formulate a background-independent quantum theory of gravity, if the most important feature of general relativity is not to be lost. If this approach proves possible and physically fruitful, then I believe that the formulation of the first background-independent physical theory will rank as Einstein’s greatest achievement.

John Stachel

In the history of physics, the opinion that almost everything is known, or that the ‘theory of the world’ is at hand, has often been widespread; after the work of Newton and after Maxwell, for example. Clearly, this was always nonsense. Personally, I find the idea that we are near to the final ‘theory of everything’ close to fantasy.

The big theoretical question is how to formulate quantum field theory consistently with what we have learned from general relativity, namely with background independence.

The tentative theories we have, such as loop gravity, strings or noncommutative geometry, are courageous attempts that are worth pursuing but badly incomplete. Loop gravity has no ambition of being a theory of everything: it is just a background-independent theory of quantum space-time. In string theory, a background-independent formulation seems as far away as ever. More crucially, none of these speculative theories has received any empirical support from experiment. Worse, phenomena such as proton decay, supersymmetric particles and signs of extra dimensions were predicted, but haven't shown up.

Rarely have we been so far from a theory of everything. Thinking that we might be close to it is the common error of those who mistake their own expectations for the ultimate truth.

Carlo Rovelli

One must distinguish a theory of fundamental forces, fields and particles that unites gravity with all other forces from a true theory of everything. The key question is what requirements of the former make it a theory of everything in the true sense — a theory that will explain both fundamental physics and how entities of genuine complexity, including human beings, can arise out of this fundamental physics.

It has been suggested that only five of the 20 parameters of the standard model determine whether complexity can arise. The problem for any underlying more fundamental theory is to explain why these five parameters lie in the part of parameter space favourable for life. The key tension is as follows: the ultimate aim of the 'theory of forces and particles' enterprise is to uniquely derive an underlying fundamental theory with no free parameters at all. However, there is no reason why any such theory will lead to physics lying in the small part of possibility space that allows life to exist. If this were to happen, then this most extraordinary coincidence (fundamental physics symmetries and variational principles inevitably leading to the existence of life) would cry out for an even more fundamental explanation. How one would attain such an explanation from within the domain of physics is completely unclear.

It may indeed be helpful if there exists no parameter-free fundamental physics theory, but rather a family of theories with varying effective parameters (perhaps because of a vast array of possible vacuum states). Then one could try to explain the anthropic coincidence as either an observational selection effect (observers only exist in universes that allow life), or a physical selection effect concerning what kinds of universes are most likely to occur, as in Lee Smolin's proposal of a quasi-darwinian evolutionary process affecting which cosmological outcome is most likely.

George Ellis

I don't like the term 'theory of everything' because it suggests that there is some theory that when discovered will immediately solve all scientific problems. Clearly there isn't any such thing. I do think that we may find what I like to call a 'final theory'; that is, a single simple theory that pushes all our explanations as far as they can go. I could be wrong about this, but we'll never know without trying to find it.

Steven Weinberg

The terminology 'theory of everything' has always worried me. There is a certain physicist's arrogance about it that suggests that knowing all the physical laws would tell us everything about the world, at least in principle. Does a physical theory of 'everything' include a theory of consciousness? Does it include a theory of morality, or of human behaviour, or of aesthetics? Even if our idea of science could be expanded to incorporate these things, would we still think of it as 'physics', or would it even be reducible to physics?

As for myself, I perhaps have enough of the physicist's arrogance about me to believe that a physical 'theory of everything' should at least contain the seeds of an explanation of the phenomenon of consciousness. It seems to me that this phenomenon is such a fundamental one that it cannot be simply an accidental concomitant

of the complexity of brain action. It must be of such sophistication that the brain is enabled to dig more deeply into the fundamental workings of the Universe than are more commonplace physical systems. And if this is so, then we are very much farther from a proper understanding of the laws of nature than most physicists seem to believe.

Indeed, irrespective of the consciousness issue, in my opinion, we are nowhere close to an accurate, purely physical theory of everything. I find it remarkable how many physicists will express the view that, despite some missing details and unifying concepts, we know virtually all we need to know to describe the fully detailed physical behaviour of systems — at least in principle. Yet, there is at least one glaring omission in present physical theory. This is how small-scale quantum processes can add up, for large and complicated systems, to the almost classical behaviour of macroscopic bodies. Indeed, it is not just an omission but an actual fundamental inconsistency, sometimes referred to as the measurement paradox (or Schrödinger's cat). In my view, until this paradox is resolved we must necessarily remain very far from a physical theory of everything — whether or not such a theory exists.

Roger Penrose

THE CONTRIBUTORS

● *Gerard 't Hooft is at the Spinoza Institute and the Institute for Theoretical Physics, Utrecht University, Leuvenlaan 4, 3584 CE Utrecht, The Netherlands. e-mail: g.thooft@phys.uu.nl*

● *Leonard Susskind is in the Department of Physics, Stanford University, Stanford, California 94305-4060, USA. e-mail: susskind@stanford.edu*

● *Edward Witten is at the Institute for Advanced Study, School of Natural Sciences, Einstein Drive, Princeton, New Jersey 08540, USA. e-mail: witten@sns.ias.edu*

● *Masataka Fukugita is at the Institute for Cosmic Ray Research, University of Tokyo, Kashiwa 2778582, Japan. e-mail: fukugita@icrr.u-tokyo.ac.jp*

● *Lisa Randall is in the Department of Physics, Harvard University, Cambridge, Massachusetts 02138, USA. e-mail: randall@physics.harvard.edu*

● *Lee Smolin is at the Perimeter Institute, Waterloo, Ontario N2L 2Y5, Canada. e-mail: lsmolin@perimeterinstitute.ca*

● *John Stachel is at the Department of Physics, Boston University, Boston, Massachusetts 02215, USA. e-mail: stachel@physics.bu.edu*

● *Carlo Rovelli at the Centre for Theoretical Physics, University of Marseille, 13288 Marseille cedex 9, France. e-mail: roveli@cpt.univ-mrs.fr*

● *George Ellis is in the Department of Mathematics, University of Cape Town, Rondebosch 7700, Cape Town, South Africa. e-mail: ellis@maths.uct.ac.za*

● *Steven Weinberg is in the Department of Physics, University of Texas, Austin, Texas 78712, USA. e-mail: weinberg@physics.utexas.edu*

● *Roger Penrose is at the Mathematical Institute, University of Oxford, Oxford OX1 3LB, UK. e-mail: rouse@maths.ox.ac.uk*

A theoretical look at the direct detection of giant planets outside the Solar System

Adam Burrows

Department of Astronomy, The University of Arizona, Tucson, Arizona, USA

Astronomy is at times a science of unexpected discovery. When it is, and if we are lucky, new intellectual territories emerge to challenge our views of the cosmos. The recent indirect detections using high-precision Doppler spectroscopy of more than 100 giant planets orbiting more than 100 nearby stars is an example of such rare serendipity. What has been learned has shaken out preconceptions, for none of the planetary systems discovered so far is like our own. The key to unlocking a planet's chemical, structural, and evolutionary secrets, however, is the direct detection of the planet's light. Because there have been as yet no confirmed detections, a theoretical analysis of such a planet's atmosphere is necessary for guiding our search.

Direct detection of an extrasolar planet requires that its dim light be separated from under the glare of its bright parent star. However, such high-contrast imaging (one part in 10^7 to one part in 10^{10} in the visible) has not so far been achieved. Instead, the vast majority of known extrasolar giant planets (EGPs) have been discovered from the ground using the indirect technique of high-precision stellar spectroscopy (refs 1, 2; see ref. 3 for an up-to-date listing). Owing to gravitational attraction, an orbiting planet induces a Doppler wobble in its parent star. If the planet is massive and close enough, the periodic variation in the stellar spectral lines can be measured. The planet's period (P), eccentricity (e), orbital semi-major axis (a), and projected mass ($M_p \sin(i)$), where i is the inclination of the orbit, can thereby be determined. The larger $M_p \sin(i)$ is, the larger is the signal. This is the reason that the first planets detected were the EGPs. Terrestrial planets, such as Earth and Venus, are ~ 300 times lighter than Jupiter, while ice giants, such as Uranus and Neptune, are ~ 20 times lighter.

Before I delve into the physical theory of EGPs and their direct detection, I summarize the basic facts of the known members of the EGPs family. The first EGP to be found was 51 Pegasi b (ref. 1) and it is in a tight 4.2-day orbit, one hundred times closer to its primary than is Jupiter to the Sun. To date, more than 140 EGPs or planets have been discovered, more than 25 of which are in more than ten multiple systems. 55 Cancri houses a quadruplet⁴, one of which has a mass near that of Neptune (~ 17 Earth masses), ν Andromedae house a triplet, and GJ 876 houses a doublet in a two-to-one orbital resonance. (We follow the convention by which the planet's name is given by the star's name, with an appended lower-case letter, either b, c or d, in discovery order).

The projected masses of the known Doppler planets vary from as little as $\sim 0.06 M_J$ to above $10 M_J$ (where M_J is a Jupiter mass, which is 318 Earth masses or roughly 10^{-3} solar masses). The more massive objects may be brown dwarfs with a different provenance (see Box 1). Radial-velocity (Doppler) techniques cannot distinguish EGPs and Neptune-mass planets from brown dwarfs. The orbital periods of the known EGPs span a vast range from ~ 1.2 days to ~ 12 years, their semi-major axes extend from ~ 0.022 to ~ 6.0 AU (where an AU is an astronomical unit, the distance between the Earth and the Sun), and their orbital eccentricities vary from 0.0 to above 0.9. For comparison, Jupiter resides 5.2 AU from our Sun, has an orbital period of ~ 12 years, and has an orbital eccentricity of ~ 0.05 . Table 1 provides these basic data for a representative subset of the known EGPs. The extremely close-in EGPs, such as 51 Peg b, τ Bootis b, HD 209458b, and OGLE-TR56b⁵⁻⁷, were a surprise, but no more so than was the heterogeneity of the

masses and orbital properties of the emerging EGP family. To be sure, the Doppler technique selects for the closer representatives, but they must exist to be detected. As would be expected from tidal dissipation, the close-in EGPs with orbital distances smaller than ~ 0.06 AU all have nearly circular orbits.

There seems to be correlation between the probability of finding an EGP and the metallicity of its parent star. The 'metallicity' of a star is the mass fraction of elements, such as carbon, oxygen, neon, magnesium, silicon and iron, that are heavier than helium. Hydrogen and helium predominate in stars and giant planets, comprising $\sim 98\%$ by mass of the Sun. The more super-solar the heavy-element composition of the potential parent, the more likely we are to find an EGP in orbit. This may be a hint concerning the processes of giant-planet formation, and is in keeping with the 3–5 times solar excesses measured in Jupiter and Saturn. The current census reveals that there is a $\sim 5\%$ *a priori* chance of finding a giant planet by the Doppler technique around a nearby (≤ 50 parsecs $\equiv 160$ light-years) star, but a $\sim 20\%$ chance of finding one around a star with at least twice the Sun's metallicity (J. Valenti and D. A. Fischer, manuscript in preparation).

Presumably, the inclinations of EGP orbits are distributed randomly on the sky. Hence, the probability that the orbit is edge-on ($i = 90^\circ$) is approximately $R_*/(2a)$, where R_* is the stellar radius. Given this, the close-in EGPs have the largest chance of transiting the stellar disk, during which time the star will dim by a fraction $(R_p/R_*)^2$, where R_p is the planet's radius. Because R_J (the radius of Jupiter, $\sim 7.14 \times 10^4$ kilometres) is roughly 10% of the radius of the Sun, this ratio is expected to be roughly 1%. A 1% dimming is easily

Box 1 Brown dwarfs

Brown dwarfs are substellar-mass objects (≤ 0.07 solar masses $\equiv \sim 75 M_J$) that are unable to ignite light hydrogen stably to become a star, but are otherwise formed like stars. The radiative surface losses of a star balance the thermonuclear power generated in its core. This requires sufficient mass. The surface losses of a less-massive brown dwarf are not fully compensated by thermonuclear burning and it cools inexorably after formation over a Hubble time. Nevertheless, brown dwarfs constitute the low-mass, low-temperature extension of the stellar family and are an important subject in their own right^{29,70}. Masses in the range of $\sim 10 M_J$ to $\sim 75 M_J$ are frequently discussed, but overlap with the mass distribution of the EGP family is certainly possible.

detectable from the ground. At $a = 0.045$ AU and a distance (d) of 47 pc, the planet around the F8V/G0V star HD 209458 was the first of only a handful of EGPs that are now known to transit their primaries and a periodic dimming at the $\sim 1.6\%$ level was measured^{8–10}. The transit HD 209458 lasts ~ 3 h (out of a total period of 3.524738 days). This was followed by the photometrically selected transiting EGPs OGLE-TR-56b, OGLE-TR-113b, OGLE-TR-132b, OGLE-TR-111b, and TrES-1^{5–7,11–13}. Many more EGP transits are anticipated during the Kepler¹⁴ and Corot¹⁵ space missions. These projects are focused on detecting transits around a fraction of the tens of thousands of stars they will monitor and will boast photometric accuracy ($\sim 10^{-5}$) sufficient to measure not only transits by EGPs, but by Earth-like planets. The import of an EGP transit lies in the simultaneous measurement of both the orbital inclination (and, hence, with Doppler spectroscopy, the mass) and the radius of the planet. Knowledge of R_p and M_p (with some knowledge of the star) can be used to constrain theories of the structure and evolution of the close-in EGP^{16–18}. At present, non-transiting EGPs are mute concerning such physical information.

HD 209458 is close and bright enough that the STIS instrument on the Hubble Space Telescope (HST) was used not only to obtain photometric precision of $\sim 0.01\%$ (ref. 10), but to distinguish a difference at the $4\text{-}\sigma$ level in the planetary transit radius in and out of the Na–D line at $0.589\text{ }\mu\text{m}$. In this way, neutral sodium atoms were discovered in HD 209458b's atmosphere^{19–21}. Though indirect, this is the first measurement of the composition of the atmosphere of an extrasolar planet. Since then, the Lyman- α line of hydrogen has similarly been detected in HD 209458b's atmosphere²², and by the large magnitude ($\sim 15\%$) of the photometric dip at this ultraviolet wavelength (λ) a planetary wind²³ comprising molecular break-up products has been inferred. However interesting, transits are rare and are no substitute for direct imaging and optical and infrared spectra. Spectra can provide diagnostics for atmospheric composition, radius, gravity and mass. Images are ground truth for the existence of a planet and provide orbital information that complements the information gleaned from Doppler measurements. Furthermore, direct detection might be able to distinguish the

different models of giant-planet formation, such as nucleation around an ice/rock core²⁴ and direct collapse²⁵, and can probe the outer orbits where the majority of EGPs might reside.

Because the indirect radial-velocity technique of EGP discovery selects for the closer variety, it is likely that a large reservoir of giants exists at distances and orbital periods beyond the reach of Doppler spectroscopy. Furthermore, the best theory for the orbits of the closest EGPs is that they migrated in from further out during the early phase of star and planet formation²⁶. This too would imply that a large pool of EGPs resides at larger separations. Indeed, it may be that the majority of stars in the solar neighbourhood harbour planetary systems that only new techniques can reveal. This is where the direct planet detection methods, most effective at large angular distances from the parent star, will come into their own.

Theoretical atmospheres and chemistry of EGPs

After formation, without any significant internal sources of energy, an EGP gradually cools and shrinks. Its rate of cooling can be moderated by stellar irradiation, or by hydrogen/helium phase separation when old and light²⁷, but is inexorable. Jupiter itself is still cooling and its total infrared plus optical luminosity is about twice the power intercepted from the Sun. The rate of cooling is a function of mass and composition, with more massive EGPs cooling more slowly. Hence, the instantaneous state of an EGP is a function of mass, age, composition, orbital distance and stellar type, not just mass and composition.

Unlike a star, EGP atmospheric temperatures are sufficiently low that chemistry is destiny. This is a distinguishing characteristic of substellar-mass objects. The atmosphere of a gaseous giant planet is the thin outer skin of molecules that regulates its emission spectrum and cooling rate. Molecular hydrogen (H_2) is the overwhelmingly dominant constituent, followed by atomic helium. An EGP's effective temperature (T_{eff} , the temperature of its 'photosphere') can vary from $\sim 1,500$ K for the more massive EGPs at birth to ~ 50 K for the least massive EGPs after a Hubble time. This wide range translates into a rich variety of atmospheric constituents that for a given mass and elemental composition evolves significantly.

Table 1 Interesting EGPs listed by angular separation

EGP	$a(1+e)/d(^{\circ})^*$	star	a (AU)	d (pc)	P	$M_p \sin(i)$ (M_J)	e
ϵ Eri b	1.61	K2V	3.3	3.2	6.85 yr	0.86	0.61
55 Cnc d	0.51	G8V	5.9	13.4	14.7 yr	4.05	0.16
47 UMa c	0.31	G0V	3.73	13.3	7.10 yr	0.76	0.1
HD 160691c	0.27	G3IV-V	2.3	15.3	3.56 yr	~ 1	~ 0.8
ν And d	0.27	F8V	2.50	13.5	3.47 yr	4.61	0.41
HD 39091b	0.26	G1IV	3.34	20.6	5.70 yr	10.3	0.62
Gl 777A b	0.23	G6V	3.65	15.9	7.15 yr	1.15	~ 0
14 Her b	0.20	K0V	2.5	17	4.51 yr	3.3	0.33
47 UMa b	0.17	G0V	2.09	13.3	2.98 yr	2.54	0.06
HD 33636b	0.17	G0V	3.56	28.7	4.43 yr	7.71	0.41
HD 10647b	0.16	F9V	2.10	17.3	2.89 yr	1.17	0.32
γ Cep b	0.15	K2V	1.8	11.8	2.5 yr	1.25	~ 0
HD 147513b	0.15	G3V	1.26	12.9	1.48 yr	1.0	0.52
HD 216437b	0.134	G4V	2.7	26.5	3.54 yr	2.1	0.34
HD 160691b	0.127	G3IV-V	1.48	15.3	1.74 yr	1.7	0.31
HD 70642b	0.121	G5IV-V	3.3	29	4.79 yr	2.0	0.10
HD 50554b	0.109	F8V	2.38	31.03	3.50 yr	4.9	0.42
HD 106252b	0.108	G0V	2.61	37.44	4.11 yr	6.81	0.54
HD 168443c	0.107	G5V	2.87	33	4.76 yr	17.1	0.23
HD 10697b	0.075	G5IV	2.0	30	2.99 yr	6.59	0.12
ν And c	0.072	F8V	0.83	13.5	241 d	2.11	0.18
GJ 876b	0.049	M4V	0.21	4.72	61.0 d	1.89	0.1
GJ 876c	0.036	M4V	0.13	4.72	30.1 d	0.56	0.27
HD 114762b	0.017	F9V	0.35	28	84.0 d	11.0	0.34
55 Cnc b	8.4×10^{-3}	G8V	0.12	13.4	14.7 d	0.84	0.02
ν And b	4.5×10^{-3}	F8V	0.059	13.5	4.62 d	0.71	0.034
51 Peg b	3.4×10^{-3}	G2V	0.05	14.7	4.23 d	0.44	0.01
τ Boo b	3.3×10^{-3}	F7V	0.05	15	3.31 d	4.09	~ 0
HD 49674b	1.6×10^{-3}	G5V	0.057	40.7	4.95 d	0.12	0.17
HD 209458b	9.6×10^{-4}	G0V	0.045	47	3.52 d	0.69	~ 0
HD 83443b	9.4×10^{-4}	K0V	0.038	43.5	2.99 d	0.35	0.08
OGLE-TR56b	1.5×10^{-5}	G0V	0.023	$\sim 1,500$	1.21 d	1.45	~ 0

*Maximum possible angular separation (at apoapse/apastron).

At birth, Jupiter had a T_{eff} near 600–1,000 K and the appearance of a T-dwarf²⁸ brown dwarf. It had no ammonia or water clouds and, owing to the presence of atomic sodium in its hot atmosphere, had a purple/magenta colour in the optical²⁹. Its atmosphere was depleted of aluminum, silicon, iron, calcium and magnesium owing to the formation and settling-to-depth of the refractory silicates ('dirt') that condense in the temperature range ~1,700–2,500 K (refs 30–32). Water vapour (steam) was the major reservoir of oxygen, gaseous methane was the major reservoir of carbon, gaseous ammonia and molecular nitrogen were the reservoirs of nitrogen, and H_2S was the reservoir of sulphur. As it cooled, the layer of alkali metals was buried below the photosphere to higher pressures, but gaseous H_2 , H_2O , NH_3 and CH_4 persisted to dominate the atmospheric composition. At a T_{eff} of ~400 K, water condensed in the upper atmosphere and water clouds appeared. This occurred within Jupiter's first 100 million years. Within less than a gigayear, when T_{eff} reached ~160 K, ammonia clouds appeared on top of the water clouds, and this layering persists now. Stellar irradiation retards cloud formation, as does a large EGP mass, which keeps the EGP hot longer. Around a G2V star like the Sun, at 5 Gyr and for an EGP mass of $1.0M_{\text{J}}$, water clouds form at 1.5 AU, whereas ammonia clouds form beyond 4.5 AU (ref. 33). Jupiter's and Saturn's current effective temperatures are 124.4 and 95 K, respectively. Jupiter's orbital distance and age are 5.2 AU and 4.6 Gyr. The orbital distance, mass, and radius of a coeval Saturn are 9.5 AU, $0.3M_{\text{J}}$, and $0.85R_{\text{J}}$.

However, as an EGP of whatever mass cools, its atmospheric composition evolves through a similar chemical and condensation sequence. Figure 1 depicts the atmospheric temperature/pressure profile for a sequence of $1-M_{\text{J}}$, 5-Gyr models as a function of orbital distance from G2V star. As the planet 'moves' outward, its atmospheric temperature at a given pressure decreases. Superposed on the plot are the H_2O and NH_3 condensation lines. In an approximate sense, a given atmospheric composition and temperature can result from many combinations of orbital distance, planet mass, stellar

type and age. This lends an added degree of complexity to the study of EGPs with which the study of stars does not need to wrestle.

The atmospheres of close-in EGPs ('roasters') at orbital distances of ~0.02–0.07 AU from a G, F or K star are heated and maintained at temperatures of 1,000–2,000 K, roughly independent of planet mass or composition. An edge star of the solar-composition, hydrogen-burning main-sequence stars ($M_{\star} \approx 75M_{\text{J}}$) has a T_{eff} of ~1,700 K. Therefore, an irradiated EGP, with a radius comparable to that of such a star, can be as luminous. Its atmospheric composition is predominantly H_2 , He, H_2O , Na, K and CO. At high temperatures, carbon is generally in carbon monoxide. This is the dominant molecule of carbon for M dwarfs with T_{eff} values of 2,200–3,500 K. At the highest T_{eff} values, clouds of iron particulates can form and persist in the upper atmosphere, as may be the case in HD 209458b. There are, however, significant day/night differences and unique reflective properties that distinguish a roaster from a lone and isolated edge star. Exotic general circulation models^{34–36} may soon be necessary to understand the equatorial currents, jet streams, day/night differences, terminator chemistry, and global wind dynamics of severely irradiated roasters in particular, and of orbiting, rotating EGPs in general.

It is useful to note that a young EGP in a wide orbit with a mass of $1.0M_{\text{J}}$ to $5.0M_{\text{J}}$ has an atmosphere and spectrum that are similar to those of an old brown dwarf with a mass of 30–60 M_{J} . As it evolves, the spectroscopic class of a giant planet can transition from that of a hot M dwarf, into an L dwarf (where the silicate clouds are within the atmosphere), then into a T dwarf, ending up in the territory, as yet unexplored, between the jovian planets and the 'stars'. If its mass is low enough, an EGP can cool within gigayears to assume the aspect of our jovian planets. Hence, by chemistry, clouds and T_{eff} , the study of brown dwarfs and EGPs are inextricably linked.

Finally, the best theoretical fits to Saturn's internal structure suggest that it contains a core of heavy elements of 5–20 Earth masses³⁷. This core of (perhaps) ice and rock may have been the nucleus around which Saturn formed, and resembles the ice giants Neptune and Uranus. Neptune and Uranus may be aborted giant planets that were able to accrete very little hydrogen from the protosolar/protoplanetary nebula. The Neptune-mass extrasolar planet 55 Can e may be a stripped or aborted EGP. An alternative mode of giant-planet formation is by direct collapse²⁵. For such a scenario, one would expect a closer correspondence between the heavy-element abundances of planet and parent star. Hence, the composition of its atmosphere and its heavy-element-dependent radius might be keys to an EGP's formation. These are, in principle, measurable.

Spectral features of EGPs

In principle, as with stellar atmospheres, direct detection of the spectrum of an extrasolar giant planet can reveal its elemental composition, radius, gravity ($GM_{\text{p}}/R_{\text{p}}^2$), and T_{eff} . Furthermore, when a cloud dwells in its atmosphere, its associated absorption and scattering properties might be used to determine the cloud's particle size and make-up. Moreover, short-term temporal variations of the planet's flux and spectrum might indicate rotation and/or meteorology. Finally, irradiation introduces the star–planet–Earth angle as an important parameter, so the orbit's orientation and instantaneous orbital phase must be factored in ('Phase functions for EGPs and orbital orientation' section). Along with the dependences on M_{p} , age, stellar type and orbital distance, this variety of influences and parameters makes the study of EGP spectral signatures and light curves, and their inversion to obtain planetary properties, rather complicated.

Nevertheless, the molecular mix described in the 'Theoretical atmospheres and chemistry of EGPs' section determines the emergent and reflected spectrum. Although H_2 is abundant, it has no permanent electric dipole moment, and, hence, a very low photon-absorption cross-section in the optical and infrared. Similarly,

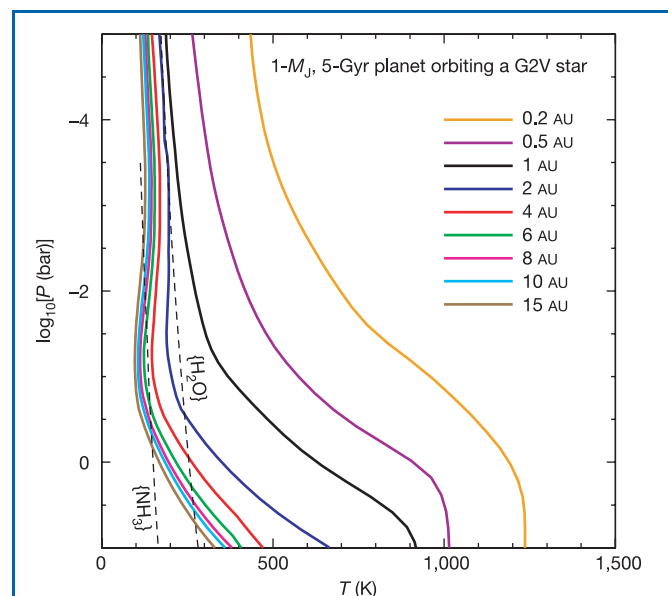


Figure 1 Profiles of atmospheric temperature (in Kelvin) versus the logarithm base ten of the pressure (in bars) for a family of irradiated $1-M_{\text{J}}$ EGPs around a G2V star as a function of orbital distance. Note that the pressure is decreasing along the ordinate, which thus resembles altitude. The orbits are assumed to be circular, the planets are assumed to have a radius of $1 R_{\text{J}}$, and the orbital separations vary from 0.2 to 15 AU. The intercepts with the dashed lines identified with either NH_3 or H_2O denote the positions where the corresponding clouds form. Taken from ref. 33. See text for a discussion.

helium is all but transparent. The result is that gaseous water vapour, with its strong absorption features from 0.94 to $\sim 7\ \mu\text{m}$, can define much of an EGP's spectrum. Because water resides in both the Earth's and an EGP's atmosphere, the water bands that bracket and determine the Earth's photometric windows at $\sim 1.0\ \mu\text{m}$ (Z), $\sim 1.25\ \mu\text{m}$ (J), $\sim 1.65\ \mu\text{m}$ (H), $\sim 2.2\ \mu\text{m}$ (K), $\sim 3.45\ \mu\text{m}$ (L'), and $\sim 4\text{--}5\ \mu\text{m}$ (M), through which ground-based infrared astronomy is possible, are exactly the same windows in an EGP or brown dwarf atmosphere through which emergent flux can pour. Thus, and fortuitously for brown dwarf observations, the emission peaks for substellar-mass objects coincide with the classic terrestrial atmospheric bandpasses.

In lieu of measurements, theory fills the vacuum. Figure 2, taken from ref. 33, depicts 'phase-averaged'³⁸ planet/star flux ratios (f) from 0.5 to $30\ \mu\text{m}$ for a $1-M_J/5\text{-Gyr}$ EGP in a circular orbit at various distances from a G2V star like the Sun. These models are the same as those depicted in Fig. 1. Similar plots for different assumed parameters can be generated. The water absorption troughs are manifest throughout. For the closer EGPs at higher atmospheric temperatures, carbon resides in CO and methane features are weak. For these close-in EGPs, the Na–D line at $0.589\ \mu\text{m}$ and the corresponding resonance line of K I at $0.77\ \mu\text{m}$ are important absorbers, suppressing flux in the visible bands. Otherwise, the optical flux is buoyed by Rayleigh scattering of stellar light. As a increases, methane forms and the methane absorption features in the optical (most of the undulations seen in Fig. 2 for $a \gtrsim 0.5\ \text{AU}$ shortward of $1\ \mu\text{m}$), at $\sim 3.3\ \mu\text{m}$, and at $\sim 7.8\ \mu\text{m}$ appear. Concomitantly, Na and K disappear from the atmosphere and the fluxes from ~ 1.5 to $\sim 4\ \mu\text{m}$ drop.

For all models, the mid-infrared fluxes longward of $\sim 4\ \mu\text{m}$ are due to self-emission, not reflection. As Fig. 2 makes clear, for larger

orbital distances a bifurcation between a reflection component in the optical and an emission component in the mid-infrared appears. This separation into components is not as straightforward for the closer, more massive, or younger family members. For these EGPs, either the large residual heat coming from the core or the severe insolation prop up the fluxes from 1 to $4\ \mu\text{m}$. The more massive EGPs, or, for a given mass, the younger EGPs, have larger J, H and K band fluxes. As a result, these bands are diagnostic of mass and age. For EGPs with large orbital distances, the wavelength range from 1.5 to $4\ \mu\text{m}$ between the reflection and emission components may be the least favourable search space, unless the substellar-mass object is massive or young.

When water or ammonia clouds form, scattering off them enhances the optical fluxes, whereas absorption by them suppresses fluxes at longer wavelengths in, for example, the $4\text{--}5\text{-}\mu\text{m}$ window. Because water and ammonia clouds form in the middle of this distance sequence, the reflection efficiency (or 'albedo'; see 'Phase functions for EGPs and orbital orientation' section) is not a monotonic function of a . These effects are incorporated into Fig. 2, but their precise magnitude depends upon unknown cloud particle size, composition and patchiness. As a consequence, direct spectral measurements might constrain cloud properties.

Importantly, trace non-equilibrium molecular species, difficult to model, can be present in quantities sufficient to alter colours. Such a 'chromophore', whose molecular nature is not yet known, absorbs in the blue and creates the reddish cast of both Jupiter and Saturn, lowering their albedos shortward of $0.55\ \mu\text{m}$ by a factor of $\sim 1.5\text{--}2$. (Chromophores were not modelled to produce Fig. 2).

As Fig. 2 suggests, the planet/star contrast ratio f is better in the mid- to far-infrared, particularly at wide separations. For such separations, the contrast ratio in the optical can sink to 10^{-10} . For the closest-in EGPs, such as HD 209458b, OGLE-TR56b, 51 Peg b, τ Boo b, the contrast ratio in the optical is between 10^{-5} and 10^{-6} and is more favourable. Such EGPs are not shown in Fig. 2; there are in fact about 20 known EGPs with orbital distances less than 0.08 AU. Owing to the possible formation of iron clouds in their atmospheres, HD 209458b and OGLE-TR56b may be brighter in the optical than 51 Peg b and may have higher reflection albedos. Figure 3 portrays a generic absolute flux spectrum at 10 pc of a close-in EGP ('Class V' in the nomenclature of ref. 38), not unlike HD 209458b. Highlighted are the positions of some of the important spectral features. Because modern telescopes can easily detect fluxes at the millijansky level, Fig. 3 demonstrates that the fluxes themselves are not small. The problem is seeing the planet from under the glare of the star (see 'Ground-based and space-based telescopes for direct detection' section). At 10 pc and an orbital separation of 0.05 AU, the maximum angular separation is a challenging ~ 5 milliarcsecs.

Phase functions for EGPs and orbital orientation

The planetary albedo and the phases executed as planets traverse their orbits are central quantities in the theory of EGP light curves. In addition, as we discuss in this section, the wavelength-dependent albedos are strong functions of orbital distance as well. Furthermore, the changing orientation of the illuminated face of a planet from the Earth's perspective translates into a light curve that can show significant flux and colour variations. In Fig. 2, these variations were averaged out over the orbit, which was assumed to be circular. The longitude independence of Jupiter's temperature–pressure profile results in little day/night variation in the mid-infrared and, hence, little phase variation, but such a planet is too cold to be self-luminous enough in the optical for its reflected component not to dominate at these shorter wavelengths. Hence, Jupiter's optical fluxes can vary from superior conjunction (full face) to first quarter (90° from superior conjunction, not seen from Earth) or last quarter (270° from superior conjunction, also not seen

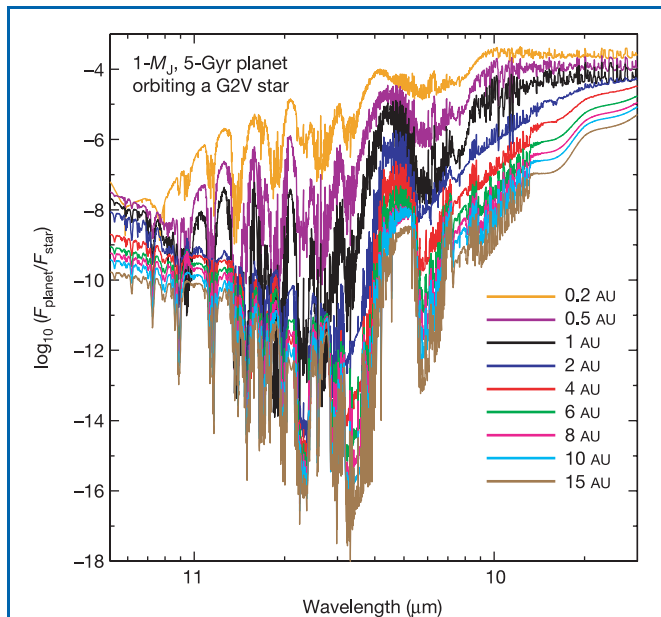


Figure 2 Planet to star flux ratios versus wavelength from 0.5 to $30\ \mu\text{m}$ for a $1-M_J$ EGP with an age of 5 Gyr orbiting a G2V main-sequence star similar to the Sun. This figure portrays ratio spectra as a function of orbital distance from 0.2 to 15 AU. Zero eccentricity has been assumed and the planet spectra have been phase-averaged as described in ref. 44. The associated temperature–pressure profiles are given in Fig. 1. Note that the planet/star flux ratio is most favourable in the mid-infrared. Water features at 0.94, 1.2, 1.4, 1.9, 2.6 and $6.5\text{--}8\ \mu\text{m}$, methane features in the optical and at 0.89, 2.2, 3.3 and $7.8\ \mu\text{m}$, carbon monoxide features at 2.3 and $4.67\ \mu\text{m}$, the Na–D doublet at $0.589\ \mu\text{m}$, and the K I doublet at $0.77\ \mu\text{m}$ help to shape these spectra. Taken from ref. 33. See text for discussion.

from Earth) by a factor of ~ 3 (ref. 39). Note that the phase dependence of an EGP's light curve in the mid-infrared will depend on the degree to which heat can be efficiently redistributed over its entire face. This will depend on three-dimensional general-circulation-model effects that have not been worked out. For the close-in EGPs, owing to expected day/night temperature differences^{40,41}, it is likely that there will be phase variations at all wavelengths. In particular, phase variations at thermal wavelengths are likely to shed light on the atmospheric dynamics and longitudinal temperature distribution of an EGP.

Because its orbit and orientation play such an important role in an EGP's flux at the Earth and in its interpretation, we summarize the basic formulae and concepts. We restrict ourselves to the optical, for which the concept of an albedo has a clear meaning, but note that the approach we summarize has general applicability.

The planet/star flux ratio (f) is given by:

$$f = p(R_p/R)^2 \Phi(\alpha) \quad (1)$$

where R is the planet–star distance, p is the geometric albedo, $\Phi(\alpha)$ is the phase function, and α is the star–EGP–Earth angle. $\Phi(\alpha)$ is normalized to be 1.0 at full face, thereby defining the geometric albedo, and is a decreasing function of α . For so-called ‘Lambert’ reflection in which an incident ray on a planetary patch emerges uniformly over the exit hemisphere, p is $2/3$ for purely scattering atmospheres and $\Phi(\alpha)$ is given by the formula:

$$\Phi(\alpha) = \frac{\sin(\alpha) + (\pi - \alpha)\cos(\alpha)}{\pi} \quad (2)$$

However, EGP atmospheres are absorbing and the anisotropy of the single scattering phase function for grains, droplets or molecules results in non-lambertian behaviour. For instance, back-scattering off cloud particles can introduce an ‘opposition’ effect for which the planet appears ‘anomalously’ bright at small α values. This spike might be a useful signature of cloud particle size. Moreover, the light scattered from EGPs is likely to be strongly polarized⁴². The degree of polarization as a function of wavelength and phase angle α can also be used to determine cloud properties. However, polarization will be rather more difficult to measure.

Both p and $\Phi(\alpha)$ are functions of wavelength, but the wavelength-

dependence of p is the most severe. In fact, for cloud-free atmospheres, owing to strong absorption by molecular bands, p can be as low as 0.03. Rayleigh scattering serves to support p , but mostly in the blue and ultraviolet, where, however, chromophores can decrease it. The presence of clouds increases p significantly. For instance, at $0.48 \mu\text{m}$, Jupiter's geometric albedo is ~ 0.46 and Saturn's is 0.39 (ref. 43). Note that for orbital distances less than 1.5 AU, we expect the atmospheres of most EGPs to be clear. The albedo would be correspondingly low. As a consequence, the theoretical albedo is very non-monotonic with distance, ranging in the visible from perhaps ~ 0.3 at 0.05 AU, to ~ 0.05 at 0.2 AU, to ~ 0.4 at 4 AU, to ~ 0.7 at 15 AU (refs 33, 38, 44). In the visible ($\sim 0.55 \mu\text{m}$), the geometric albedo for a roaster is severely suppressed by Na–D at $0.589 \mu\text{m}$. Owing to a methane feature, the geometric albedo can vary from 0.05 at $\sim 0.6 \mu\text{m}$ to ~ 0.4 at $0.625 \mu\text{m}$. Hence, variations with wavelength and with orbital distance by factors of 2 to 10 are not unexpected. Those planning programmes of direct detection should be aware of such possibilities.

$\Phi(\alpha)$ and p must be calculated or measured, but the sole dependence of $\Phi(\alpha)$ on α belies the complications introduced by an orbit's inclination angle (i), eccentricity (e), argument of periastron (ω), and longitude of ascending node (Ω). Along with the period (P) and an arbitrary zero of time, these are the so-called keplerian elements of an orbit. Figure 4 defines these orientational and orbital parameters. In the plane of the orbit, the angle between the planet and the periastron/periapse (distance of closest approach to the star) at the star is θ . In the jargon of celestial mechanics, θ is the so-called ‘true anomaly’. For an edge-on orbit ($i = 90^\circ$), and one for which the line of nodes is perpendicular to the line of sight ($\Omega = 90^\circ$) and parallel to the star–periapse line ($\omega = 0^\circ$), θ is complementary to α ($\alpha = 90^\circ - \theta$). As a result, $\theta = 0^\circ$ at $\alpha = 90^\circ$ (greatest elongation) and increases with time. Also, for such an edge-on orbit, $\alpha = 0^\circ$ at superior conjunction. In general:

$$\cos(\alpha) = \sin(\theta + \omega)\sin(i)\sin(\Omega) - \cos(\Omega)\cos(\theta + \omega) \quad (3)$$

This is merely an application of the law of cosines.

For a circular orbit, R is equal to the semi-major axis (a). However, a planet in an eccentric orbit can experience significant variation in R , and, therefore, stellar insolation (by a factor of $(1 + e)^2/(1 - e)^2$). For example, if $e = 0.3$, the stellar flux varies by

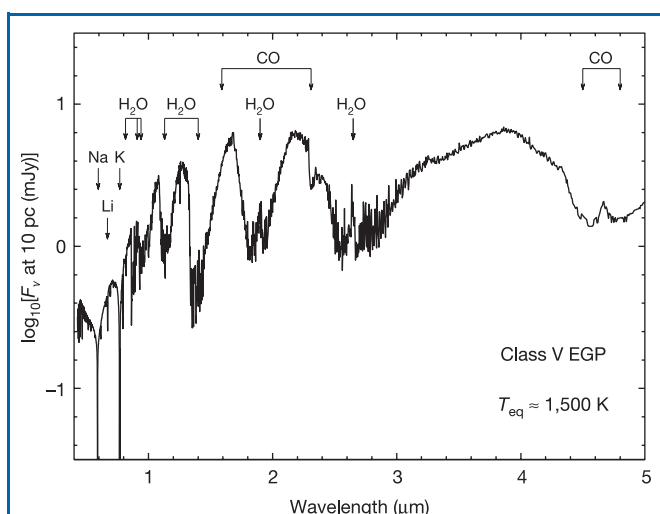


Figure 3 The logarithm of the absolute flux in millijanskys ($\equiv 10^{-26} \text{ ergs cm}^{-2} \text{ s}^{-1} \text{ Hz}^{-1}$) at 10 pc for a ‘Class V’ roaster versus wavelength from 0.4 to $5 \mu\text{m}$. This could be a $1-M_J$ EGP in a 0.05-AU orbit around a solar-type star. The planet spectrum has been phase-averaged as described in refs 38 and 44. The positions of various relevant molecular bands and atomic lines are shown. Figure is taken from ref. 44. See text for discussion.

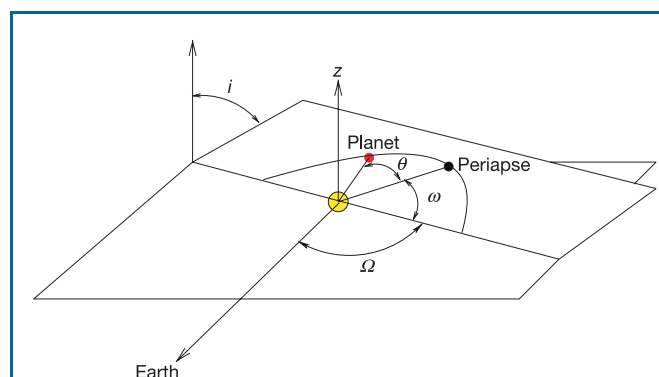


Figure 4 Keplerian orbital elements. The intersection of the orbit plane with the observational plane defines the angles i , Ω , ω and θ . The angle between the observer (Earth) and the line of nodes (intersection of the orbit plane with the horizontal plane) is the longitude of the ascending node (Ω), the angle between the line of nodes and the focus (star, in yellow) to periapse (black dot) line is the argument of periastron (ω), the angle between the orbit plane and the z axis (perpendicular to the horizontal plane) is the inclination (i), and the angle between the focus–periapse line and the position of the planet (red dot) is the true anomaly (θ). See text for details.

~3.5 along its orbit. For $e = 0.6$, this variation is a factor of 16! Such eccentricities are by no means rare in the sample of known EGPs (see Table 1). Therefore, it is possible for the composition of an EGP atmosphere to change significantly during its orbit, for clouds to appear and disappear, and for there to be delays ('hysteresis') in the accommodation of a planet's atmosphere to a varying 'insolation' regime. Ignoring the latter, equations (1) and (3) can be combined with $\Phi(\alpha)$ and the standard keplerian formula connecting θ and time for an orbit with a given P and e to derive an EGP's light curve as a function of wavelength, i , e , Ω , ω and time. The upshot is that, depending upon orientation and eccentricity, the brightness of an EGP can vary in its orbit not at all (for a face-on EGP in a circular orbit) or quite dramatically (such as for highly eccentric orbits at high inclination angles). Astrometric measurements of stellar wobble induced by EGPs can yield the entire orbit (including inclination), so data from the Space Interferometry Mission (SIM)⁴⁵ (expected to achieve 1-microarcsecond narrow-angle accuracy) or Gaia⁴⁶ could provide important supplementary data to aid in the interpretation of direct detections of EGPs.

As Saturn itself demonstrates, depending upon orbital orientation, planetary rings can greatly augment reflected light^{39,47}. Their possible presence is a wild card in the interpretation of direct EGP signatures. Also, since $\Phi(\alpha)$ is wavelength-dependent, the potentially large variation in reflected optical flux with epoch will be complemented by an interesting variation in colour. The phase functions $\Phi(\alpha)$ are wavelength-dependent. For example, planets should execute trajectories in the colour-colour space $V-R$ versus $B-V$, where B , V and R are the standard blue, visible and red bands. These trajectories will be functions of cloud particle size, among other things, and will be useful atmospheric diagnostics. Similar behaviour in the near- and mid-infrared colours, though more modest, may be seen.

Ground-based and space-based telescopes for direct detection

As Fig. 2 implies, the wide range of planet/star contrast ratios and spectral diagnostics suggests different technological solutions to direct detection. Furthermore, the relative merits of searching in the optical, near-infrared or the mid-infrared have yet to be determined. Both ground-based (less expensive) and space-based (more capable) paths are being pursued and although a discussion that does justice to the many initiatives whose goal is the remote sensing of EGPs is far beyond the scope of this review, I summarize a few representative approaches here.

EGPs, especially if they are young, massive, and close, are bright enough that current 8–10-m-class ground-based telescopes or near-term space telescopes (such as the James Webb Space Telescope (JWST)⁴⁸ with a 6-m aperture) might be sensitive enough to pick up their light. This is particularly true in the near- and mid-infrared (see Figs 2 and 3). However, under the extreme glare of its parent star and at small angular separations in the range from a few milliarcseconds to around an arcsecond (Table 1), traditional telescope optics spills far too much light in the vicinity of the planet. The major culprits on the ground are the turbulence of the atmosphere ('seeing' and scintillation), scattering off dust and the spider mount of the secondary, and imperfections in the mirror(s). Also at issue is the stability of the optical system. Even for perfect optics, the diffraction pattern due to the finite telescope aperture leaves a characteristic 'Airy' pattern that for a jovian-type planet at 10 pc around a solar-type star would in the optical be hundreds of times brighter.

Hence, for large (8–10-m) ground-based telescopes such as the two Kecks^{49,50}, the four VLTs^{51,52}, the two Geminis⁵³, Subaru⁵⁴, the binocular LBT^{55,56} and mammoth proposed telescopes (such as the 100-m OWL⁵⁷, the 20-m GMT⁵⁸ and the 30-m GSMT⁵⁹) special efforts will be required. These include adaptive optics to compensate for atmospheric fluctuations (and mirror imperfections) with many hundreds or thousands of fast (millisecond) actuators and

very accurate wavefront sensing. The latter can, in principle, be achieved using artificial laser guide stars or stars in the field of view. (With adaptive optics, there is usually a bright star close enough to obviate the need for an artificial beacon.) Interferometry to null out the stellar light is also being pursued by the LBT, VLT and Keck, and the depth of the null is crucial, as is the angular region over which a sufficient null can be achieved. Finally, apodizing masks and/or coronagraphic spots to occult the star and, by diffractive interference, redistribute the star's light away from the planet are highly desirable (and may be necessary). Note that Dome C in Antarctica has some of the best seeing on Earth and the quietest atmosphere, so placing a giant next-generation telescope there may have its advantages⁶⁰.

An EGP-imaging system will be judged by the planet/star contrast ratio, f , it can achieve at a given wavelength and for a given angular separation from the star. Angles of 0.05 to 2.0 arcsec are contemplated (Table 1), with the requisite contrasts at smaller angles deemed too difficult for first-generation imaging. Note that the actual requirements are a function of distance to the star. As Fig. 2 indicates for a 5-Gyr/1- M_J EGP at 10 pc, f values better than 10^{-4} at 10 μm and better than 10^{-8} in the optical might be necessary. At the 4–5- μm bump, f values from 10^{-5} to 10^{-8} may be called for. Fortunately, these performance goals can be relaxed for more massive and younger EGPs at angular separations greater than $\sim 0.1''$. For the roasters, almost independently of mass and age, f reaches 10^{-3} in the mid-infrared and $\lesssim 10^{-5}$ in the optical, but at the corresponding milliarcsecond separations even these contrast ratios may be too challenging for imaging.

Figure 5 compares the theoretically required contrast ratios for the fiducial 5-Gyr/1- M_J EGP at various wavelengths (taken from Fig. 2) as a function of angular separation from a solar-type star at 10 pc with the putative capabilities of a sample of proposed imaging systems, both on the ground and in space. Contrast ratios for a 0.5-Gyr/7- M_J EGP in the H band are also shown. Orbit and orientation effects have been ignored and large error bars should be assigned to both theory and projected capability. In addition, care should be taken to compare theoretical numbers with experimental hopes for the same wavebands. In the interests of brevity, we have included multiple wavebands on Fig. 5.

Telescopes are stages of components (primary mirror, secondary mirror, lens, apertures, apodizing masks, coronagraphs, and so on) that act in series on the incident source wavefront to focus light of a desired character on instruments. Daisy-chained together, each 'optical' component convolves itself with an input wavefront to produce an output wavefront. In spatial frequency space, the operation of each stage along the optical path is to multiply the Fourier transform of the incident wavefront by a Fourier transform characteristic of that component's optical properties and geometry. If you can introduce components in the optical path that filter or alter the frequency distribution of the wavefront in such a way that its inverse (the spatial distribution of the light in the image plane) has little or no light in a two-dimensional angular realm around the star where a planet might reside, then you have a planet-imaging system. Although a telescope's classical angular resolution ($\sim \lambda/D$, where D is the diameter of the telescope primary) improves with decreasing λ , the negative effects of the atmosphere actually diminish with increasing λ . As a result, many ground-based planet-finding initiatives (MMT(AO)⁵⁶, LBT-I, VLT-I, VLT-PF, Keck-I, Gemini-XAOPI/ExAOC, OWL) are planning to optimize in the near- or mid-infrared. Figure 5 summarizes the performance goals of some of them.

In coronagraphic mode, the space-borne JWST may achieve f values of 10^{-3} to 10^{-6} for wavelengths from ~ 1.0 to $\sim 5.0 \mu\text{m}$. HST/NICMOS has already achieved comparable f values in the H band at angular separations from 0.3 to 1.0 arcsec (ref. 61). However, for single space telescopes without the atmosphere with which to contend, the optical is clearly preferred (small λ/D). Curiously,

above the atmosphere mirror imperfections and thermal flexure are still problems and an adaptive optics system is necessary to cancel the wavefront errors introduced by the corrugations that remain on an otherwise almost perfect mirror surface after state-of-the-art machining and polishing. Two major space-based projects to image EGPs are being proposed. The first, EPIC⁶², is a nulling coronagraph which converts a single telescope pupil into a multi-beam nulling interferometer, producing a null which is then filtered by an array of single-mode fibres to suppress the residual scattered light. The design goal of EPIC is for f values of 10^{-9} to 10^{-10} . The second, ECLIPSE^{63,64}, is an off-axis coronagraph with an exquisitely configured 1.8-m primary that is designed to achieve f values in the V band of better than 10^{-9} for angular separations from ~ 0.1 to ~ 2.0 arcsec. Both ECLIPSE and EPIC will be challenging, but if successful will directly detect within about seven years many EGPs in the solar neighbourhood out of 10–15 pc.

However, the flagship of the NASA 'Origins' programme, the Terrestrial Planet Finder (TPF⁶⁵), whose goal is to image planetary systems and extrasolar Earth-like planets and to obtain low-resolution ($\lambda/\Delta\lambda = 10\text{--}20$) spectra that may reveal in rudimentary fashion the O_2 , H_2O , CH_4 , O_3 or CO_2 signatures of life, will also be a formidable instrument for directly detecting and characterizing EGPs. As Fig. 5 indicates, for angular separations between ~ 0.05 and ~ 2.0 arcsec, either the more-straightforward optical coronagraphic design (TPF-C) or the multi-telescope infrared (5–20 μm) interferometer (TPF-I/Darwin) would detect EGPs much more

readily than the extrasolar Earth-like planets that are its primary targets.

For the close-in EGPs, direct imaging seems to be out of the question for the foreseeable future. However, this does not mean that the planetary flux cannot be measured. From the ground, there are a variety of techniques to use the light of the planet + star to distinguish the planetary component (particularly for known roasters). These include: (1) using precision photometry down to one part in 10^5 to measure the phase variations of the summed optical or near-infrared light; (2) measuring the motion of the light centroid, perhaps best done in the mid-infrared with an Antarctic 30-m telescope; (3) spectral deconvolution of a known EGP/star system using its RV-measured velocities and ephemeris; and (4) multi-frequency differential interferometric imaging (pioneered for Keck, among others). Further transit studies of HD 209458b (such as led to the discovery of the Na–D and Lyman- α features) are certainly warranted. The ground-based methods will be challenging, but less expensive than space-based efforts. However, there is at present in space a micro-satellite, MOST⁶⁶, with a 15-cm aperture, that is designed to achieve photometric accuracy in the optical of a few times 10^{-6} . It has on its current observing manifest programmes to stare at 51 Peg b, τ Boo b and HD 209458b. There have as yet been no announcements.

The future

With many programmes of direct planet detection planned on a

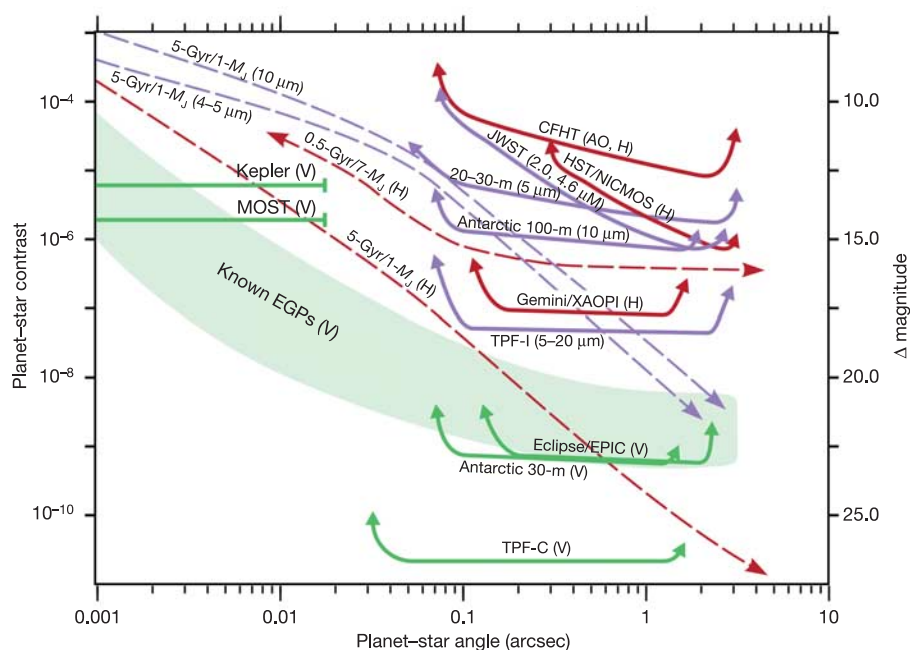


Figure 5 A comparison of the planet/star contrast ratios (and contrast magnitudes = $-2.5 \log(f)$) versus angular separation (in arcsec) achievable for some proposed planet imaging systems. A distance of 10 pc is assumed. Integration times and signal-to-noise ratios assumed vary and are taken from preliminary studies by the associated instrument teams. At H band (red), the imaging telescopes represented include the Canada/France/Hawaii Telescope (CFHT)⁶⁹, HST/NICMOS and the Gemini/XAOPI. Not shown on this plot is the MMT (AO) system, which should achieve, at a wavelength of 5 μm , f values from 10^{-4} to a few $\times 10^{-6}$ for angular separations of 0.3'' to 1.0'', respectively. The LBT (also not shown) should achieve, at a wavelength of 10 μm , an approximately ten times better performance than this. At 5 μm , a notional curve for a 20- or 30-m telescope in Antarctica and the JWST (in fact, at 4.6 μm) are provided. At 10 μm , a notional curve for a 100-m telescope in Antarctica is given. Also included on this plot is the interferometric version of TPF

(TPF-I/Darwin), which might have a sensitivity of one part in 10^7 from 5 to 20 μm . All the mid-infrared curves are in purple. In the optical (green, V), putative sensitivities for the EPIC, ECLIPSE and TPF-C instruments are plotted. Superposed are corresponding 'phase-averaged' theoretical curves (dashed) for a 5-Gyr/1- M_J EGP around a G2V star in the H band ($\sim 1.65 \mu\text{m}$), in the 4–5- μm band, and at 10 μm (see Fig. 2). Also included are a theoretical curve in the H band (dashed red) for a 0.5-Gyr/7- M_J EGP around a G2V star and a green swathe where the known EGPs may reside in the optical (V band). Note that the theoretical curves for more massive and younger EGPs than represented on this plot can be considerably higher. Orientation effects have been ignored. Each curve is for a given wavelength or bandpass and the imager and theory must be compared at the same wavelength. Very generous error bars should be assumed. The photometric sensitivity curves for MOST and Kepler are also superposed. See text for details.

large subset of the LBT, VLT, Keck, Gemini, GMT, GSMT/CELT and OWL on the ground and HST, Corot, Kepler, MOST, SIM, Gaia, TPF-C, TPF-I/Darwin, ECLIPSE, EPIC and JWST in space, during the next twenty years there will be an increasing crescendo of new results on extrasolar planets that will completely transform our view of the nature of planetary systems.

EGPs, being brighter, are the natural technological and scientific stepping stones on the path to imaging extrasolar Earths. We will encounter them first. Both the NASA and ESA roadmaps⁶⁷ have given planet detection pride of place. Strategies are now being formulated to establish a logical sequence of missions and telescope construction that will optimize the pace of discovery. Moreover, theoretical work in support of mission planning is maturing to the point where it may be ready to interpret what we observe. However, a theorist's prejudices aside, I cannot help but wonder: As the hunt for worlds beyond our Solar System quickens, an ancient curiosity stirs to ask: what will our generation see from that fabled peak in Darien⁶⁸? □

doi:10.1038/nature03244.

- Mayor, M. & Queloz, D. A Jupiter-mass companion to a solar-type star. *Nature* **378**, 355 (1995).
- Marcy, G. W. & Butler, R. P. A planetary companion to 70 Virginis. *Astrophys. J.* **464**, L147–L151 (1996).
- Schneider, J. *Extrasolar Planet Encyclopaedia* (<http://www.obspm.fr/encycl/encycl.html>) (2004).
- MacArthur, B. *et al.* Detection of a Neptune-mass planet in the ρ^1 Cancri system using the Hobby-Eberly Telescope (55 Cancri e). *Astrophys. J.* (submitted).
- Konacki, M., Torres, G., Jha, S. & Sasselov, D. An extrasolar planet that transits the disk of its parent star. *Nature* **421**, 507–509 (2003).
- Torres, G., Konacki, M., Sasselov, D. & Jha, S. New data and improved parameters for the extrasolar transiting planet OGLE-TR-56b. *Astrophys. J.* **609**, 1071–1075 (2003).
- Sasselov, D. The new transiting planet OGLE-TR-56b: orbit and atmosphere. *Astrophys. J.* **596**, 1327–1331 (2003).
- Henry, G., Marcy, G. W., Butler, R. P. & Vogt, S. S. A transiting “51 Peg-like” planet. *Astrophys. J.* **529**, L41–L44 (2000).
- Charbonneau, D., Brown, T. M., Latham, D. W. & Mayor, M. Detection of planetary transits across a Sun-like star. *Astrophys. J. Lett.* **529**, L45–L48 (2000).
- Brown, T. M., Charbonneau, D., Gilliland, R. L., Noyes, R. W. & Burrows, A. Hubble Space Telescope time-series photometry of the transiting planet of HD 209458. *Astrophys. J.* **552**, 699–709 (2001).
- Bouchy, F. *et al.* Two new “very how Jupiters” among the OGLE transiting candidates. *Astron. Astrophys.* **421**, L13–L16 (2004).
- Alonso, R. *et al.* TrES-1: The transiting planet of a bright K0V star. *Astrophys. J.* (in the press).
- Pont, F. *et al.* The “missing link”: A 4-day period transiting exoplanet around OGLE-TR-111. *Astron. Astrophys.* (2004) (in the press); preprint at (<http://arXiv.org/astro-ph/0408499>).
- Koch, D. *et al.* in *Space Telescopes and Instruments V SPIE Conf.* **3356**, 599–607 (1998).
- Antonello, E. & Ruiz, S. M. The Corot mission. *Mem. Soc. Astron. Ital.* **73**, 1241–1242 (2002).
- Burrows, A., Sudarsky, D. & Hubbard, W. B. A theory for the radius of the transiting giant planet HD 209458b. *Astrophys. J.* **594**, 545–551 (2003).
- Burrows, A., Hubeny, I., Hubbard, W. B. & Sudarsky, D. Theoretical radii of transiting giant planets: The case of OGLE-TR-56b. *Astrophys. J.* **610**, L53–L56 (2004).
- Baraffe, I., Chabrier, G., Allard, F. & Hauschildt, P. H. Evolutionary models for cool brown dwarfs and extrasolar giant planets. The case of HD 209458. *Astron. Astrophys.* **402**, 701–712 (2003).
- Charbonneau, D., Brown, T. M., Noyes, R. W. & Gilliland, R. L. Detection of an extrasolar planet atmosphere. *Astrophys. J.* **586**, 377–384 (2002).
- Seager, S. & Sasselov, D. D. Theoretical transmission spectra during extrasolar giant planet transits. *Astrophys. J.* **537**, 916–921 (2000).
- Hubbard, W. B. *et al.* Theory of extrasolar giant planet transits. *Astrophys. J.* **560**, 413–419 (2001).
- Vidal-Madjar, A. *et al.* An extended upper atmosphere around the extrasolar planet HD209458b. *Nature* **422**, 143–146 (2003).
- Burrows, A. & Lunine, J. I. Extrasolar planets—astronomical questions of origin and survival. *Nature* **378**, 333 (1995).
- Mizuno, H. Formation of the giant planets. *Prog. Theor. Phys.* **64**, 544–557 (1980).
- Boss, A. P. Gas giant protoplanet formation: Disk instability models with thermodynamics and radiative transfer. *Astrophys. J.* **563**, 367–373 (2001).
- Trilling, D. *et al.* Orbital migration of giant planets: Modeling extrasolar planets. *Astrophys. J.* **500**, 428–439 (1998).
- Fortney, J. J. & Hubbard, W. B. Effects of helium phase separation on the evolution of extrasolar giant planets. *Astrophys. J.* **608**, 1039–1049 (2004).
- Burgasser, A. *et al.* The spectra of T dwarfs. I. Near-infrared data and spectral classification. *Astrophys. J.* **564**, 421–451 (2002).
- Burrows, A., Hubbard, W. B., Lunine, J. I. & Liebert, J. The theory of brown dwarfs and extrasolar giant planets. *Rev. Mod. Phys.* **73**, 719–765 (2001).
- Burrows, A. & Sharp, C. M. Chemical equilibrium abundances in brown dwarf and extrasolar giant planet atmospheres. *Astrophys. J.* **512**, 843–863 (1999).
- Lodders, K. Alkali element chemistry in cool dwarf atmospheres. *Astrophys. J.* **519**, 793–801 (1999).
- Lodders, K. & Fegley, B. Atmospheric chemistry in giant planets, brown dwarfs, and low-mass dwarf stars. I. Carbon, nitrogen, and oxygen. *Icarus* **155**, 393–424 (2002).
- Burrows, A., Sudarsky, D. & Hubeny, I. Spectra and diagnostics for the direct detection of wide-separation extrasolar giant planets. *Astrophys. J.* **609**, 407–416 (2004).
- Menou, K., Cho, J. Y.-K., Seager, S. & Hansen, B. M. S. “Weather” variability of close-in extrasolar giant planets. *Astrophys. J.* **587**, L113–L116 (2003).
- Cho, J. Y.-K., Menou, K., Hansen, B. M. S. & Seager, S. The changing face of the extrasolar giant planet

- HD 209458b. *Astrophys. J.* **587**, L117–L120 (2003).
- Burkert, A., Lin, D. N. C., Bodenheimer, P., Jones, C. & Yorke, H. On the surface heating of synchronously-spinning short-period jovian planets. Preprint at (<http://arXiv.org/astro-ph/0312476>) (2004).
- Saumon, D. & Guillot, T. Shock compression of deuterium and the interiors of Jupiter and Saturn. *Astrophys. J.* **609**, 1170–1180 (2004).
- Sudarsky, D., Burrows, A. & Pinto, P. Albedo and reflection spectra of extrasolar giant planets. *Astrophys. J.* **538**, 885–903 (2000).
- Dyudina, U. A. *et al.* Phase light curves for extrasolar Jupiters and Saturns. Preprint at (<http://arXiv.org/astro-ph/0406390>) (2004).
- Guillot, T. & Showman, A. P. Evolution of “51 Pegasus b-like” planets. *Astron. Astrophys.* **385**, 156–165 (2002).
- Showman, A. P. & Guillot, T. Atmospheric circulation and tides of “51 Pegasus b-like” planets. *Astron. Astrophys.* **385**, 166–180 (2002).
- Seager, S., Whitney, B. A. & Sasselov, D. D. Photometric light curves and polarization of close-in extrasolar giant planets. *Astrophys. J.* **540**, 504–520 (2000).
- Karkoschka, E. Methane, ammonia, and temperature measurements of the jovian planets and Titan from CCD-spectrophotometry. *Icarus* **133**, 134–146 (1999).
- Sudarsky, D., Burrows, A. & Hubeny, I. Theoretical spectra and atmospheres of extrasolar giant planets. *Astrophys. J.* **588**, 1121–1148 (2003).
- Unwin, S. C. & Shao, M. in *Interferometry in Optical Astronomy* (eds Lena, P. J. & Quirrenbach, A.) 754–761, (2000).
- Perryman, M. A. C. in *Gaia Spectroscopy: Science and Technology* (ed. Munari, U.) *ASP Conf. Ser.* **298**, 3–4 (2003).
- Arnold, I. & Schneider, J. The detectability of extrasolar planet surroundings. I. Reflected-light photometry of unresolved rings. Preprint at (<http://arXiv.org/astro-ph/0403330>) (2004).
- Mather, J. C. & Stockman, H. S. *The Next Generation Space Telescope. Proc. SPIE* **4013**, 2–16 (2000).
- Akeson, R. L., Swain, M. R. & Colavita, M. M. in *Interferometry in Optical Astronomy* (ed. Lena, P. J.) *Proc. SPIE* **4006**, 321–327 (2000).
- Akeson, R. L. & Swain, M. R. in *From Giant Planets to Cool Stars* (eds Griffith, C. A. & Marley, M. S.) *ASP Conf. Ser.* **212**, 300 (2000).
- Paresce, F. Scientific objectives of the VLTI interferometer. *Messenger* **104**, 5–7 (2001).
- Beuzit, J.-L. *et al.* in *SF2A-2004: Semaine de l'Astrophysique Française* (eds Combes, F., Barret, D., Contini, T., Meynadier, F. & Paganii, L.) 32 (EdP-Sciences Conf. Ser., Paris, 2004).
- Macintosh, B. *et al.* in *Techniques and Instrumentation for Detection of Exoplanets* (ed. Coulter, D. R.) *Proc. SPIE* **5170**, 272–282 (2003).
- Tamura, I. & Oasa, Y. in *Proc. IAU 8th Asian-Pacific Regional Meeting: Vol. I* (eds Ikeuchi, S., Hearnshaw, J. & Hanawa, T.) 73–76 (Astron. Soc. Pac. Vol. 289, San Francisco, 2003).
- Hinz, P. M., Angel, J. R. P., Woolf, N. J., Hoffman, W. F. & McCarthy, D. W. in *Interferometry in Optical Astronomy* (ed. Lena, P. J.) *Proc. SPIE* **4006**, 349 (2000).
- Hinz, P. M. *Nulling Interferometry for Studying Other Planetary Systems: Techniques and Observations*. PhD thesis, Univ. Arizona (2001).
- Howard, T. G., Gilmozzi, R. & Hainaut, O. in *Scientific Frontiers in Research on Extrasolar Planets* (eds Deming, D. & Seager, S.) 581–586 (ASP Conf. Ser. Vol. 294, Astron. Soc. Pac., San Francisco, 2003).
- Davison, W. B., Woolf, N. J. & Angel, J. R. P. in *Future Giant Telescopes* (eds Angel, J. R. P. & Gilmozzi, R.) *Proc. SPIE* **4840**, 533–540 (2003).
- Strom, S. Toward a Giant Segmented Mirror Telescope: A Progress Report from AURA's New Initiatives Office. *BAAS* **203**, no. 24.04 (2003).
- Aristidi, E. *et al.* Antarctic site testing: First daytime seeing monitoring at Dome C. *Astron. Astrophys.* **406**, L19–L22 (2003).
- Schneider, G. *et al.* Domains of observability in the near-infrared with HST/NICMOS and adaptive-optics augmented large ground-based telescopes. Preprint at (http://nicmosis.as.arizona.edu:8000/REPORTS/NICMOS_AO_WHITEPAPER.html) (2002).
- Shao, M., Levine, B. M., Wallace, J. K., Serabyn, E. & Liu, D. T. The visible nulling coronagraph—progress towards mission and technology development. *BAAS* **203**, no. 03.04 (2003).
- Trauger, J. *et al.* Eclipse, a direct imaging investigation of nearby planetary systems. *BAAS* **197**, no. 49.07, 1486 (2000).
- Trauger, J., Hull, A. B. & Redding, D. A. Eclipse test bed for very high contrast space astronomy. *BAAS* **199**, no. 86.04, 1431 (2001).
- Beichman, C. A., Coulter, D. R., Lindensmith, C. & Lawson, P. R. in *Future Research Direction and Visions for Astronomy* (ed. Dressler, A. M.) *Proc. SPIE* **4835**, 115–121 (2002).
- Matthews, J. M., Kuschnig, R. & Shkolnik, E. in *Proc. SOHO 10/GONG 2000 Workshop: Helio- and Asteroseismology at the Dawn of the Millennium* (ed. Wilson, A.) 385–390 (ESA SP-464, ESA Publications Division, Noordwijk, 2001).
- Roadmap for the Office of Science Origins Theme*. NAS/JPL-400-1060 (NASA 2003 Strategic Plan, Document no. NP-2003-01-298-HQ, 2003).
- Keats, J. On *First Looking into Chapman's Homer in The Poetical Works of John Keats* (Macmillan, London, 1884).
- Racine, R., Walker, G. A. H., Nadeau, D., Doyon, R. & Marois, C. Speckle noise and the detection of faint companions. *Publ. Astron. Soc. Pac.* **111**, 587–594 (1999).
- Burrows, A. *et al.* A nongray theory of extrasolar giant planets and brown dwarfs. *Astrophys. J.* **491**, 856–875 (1997).

Acknowledgements I thank D. Sudarsky, I. Hubeny, W. Hubbard, J. Lunine, J. Liebert, J. Young, J. Trauger, J. Fortney, A. Li, C. Sharp, and D. Milsom for conversations or technical help during the course of this work. I thank NASA and the NASA Astrobiology Institute for their financial support.

Competing interests statement The authors declare that they have no competing financial interests.

Correspondence and requests for materials should be addressed to the author (burrows@as.arizona.edu).

Structure of human follicle-stimulating hormone in complex with its receptor

Qing R. Fan¹ & Wayne A. Hendrickson^{1,2}

¹Department of Biochemistry and Molecular Biophysics and, ²Howard Hughes Medical Institute, Columbia University, New York, New York 10032, USA

Follicle-stimulating hormone (FSH) is central to reproduction in mammals. It acts through a G-protein-coupled receptor on the surface of target cells to stimulate testicular and ovarian functions. We present here the 2.9-Å-resolution structure of a partially deglycosylated complex of human FSH bound to the extracellular hormone-binding domain of its receptor (FSHR_{HB}). The hormone is bound in a hand-clasp fashion to an elongated, curved receptor. The buried interface of the complex is large (2,600 Å²) and has a high charge density. Our analysis suggests that all glycoprotein hormones bind to their receptors in this mode and that binding specificity is mediated by key interaction sites involving both the common α - and hormone-specific β -subunits. On binding, FSH undergoes a concerted conformational change that affects protruding loops implicated in receptor activation. The FSH-FSHR_{HB} complexes form dimers in the crystal and at high concentrations in solution. Such dimers may participate in transmembrane signal transduction.

FSH is secreted from the pituitary gland to regulate reproduction in mammals^{1,2}. In females, FSH targets a receptor (FSHR) expressed only on granulosa cells, and induces the maturation of ovarian follicles^{1,3}. The levels of both FSH and FSHR rise until the middle of oestrus, and then fall precipitously as luteinizing hormone (LH) triggers ovulation. In males, FSH stimulates sertoli cell proliferation in testes and supports spermatogenesis^{1,3}. FSH is used clinically to treat anovulatory women and infertile men, and we expect that a molecular understanding of the FSH interaction with FSHR may assist in the design of contraceptive antagonists and alternative agonists.

FSH is a member of the glycoprotein hormone family, which also includes LH, chorionic gonadotropin (CG) and thyroid-stimulating hormone (TSH)⁴. These glycoprotein hormones are disulphide-rich heterodimers consisting of non-covalently associated α - and β -subunits⁴. In a given species, they share a common α -chain but derive functional specificity from different β -chains⁴. Crystal structures of human CG (hCG)^{5,6} and FSH⁷ both reveal elongated molecules with similar folds for the α - and β -chains, and a cysteine-knot motif in the central core of each subunit. The interactions between the glycoprotein hormones and their corresponding receptors are highly selective, with very few cases of cross activity³.

The glycoprotein hormones act through specific G-protein-coupled receptors (GPCRs) on target cell surfaces^{1,8,9}. FSH and TSH bind to FSHR and TSHR, respectively; LH and CG both bind to the same receptor, LHR. These receptors belong to a subfamily of GPCRs called leucine-rich-repeat-containing GPCRs (LGRs), which are characterized by an extracellular domain with multiple leucine-rich repeats (LRRs) and a rhodopsin-like domain of seven transmembrane (7TM) helices¹⁰. Unlike most other GPCR proteins, which have relatively short extracellular segments and bind small molecule ligands, the LGRs have large ectodomains and bind protein ligands including glycoprotein hormones and relaxin¹⁰.

Ligand binding and receptor activation appear to be distinct activities of glycoprotein hormone receptors. The extracellular domains of glycoprotein hormone receptors are responsible for the specificity and high affinity of ligand binding (typically at sub-nanomolar dissociation constant (K_d) values^{2,8}), whereas the transmembrane domains are responsible for receptor activation and signal transduction through G proteins^{1,8,9}. Hormone binding to the ectodomain of a glycoprotein hormone receptor somehow prompts changes in the 7TM domain that propagate across the plasma

membrane to elicit guanine nucleotide exchange in a heterotrimeric G_s protein. Activation of adenylyl cyclase for the production of cAMP ensues, thereby initiating a signalling cascade that leads to steroid synthesis^{1,8,9}. In an effort to understand ligand binding and receptor stimulation in this signalling pathway, we have determined the crystal structure of human FSH bound to the hormone-binding domain of its receptor. Our finding of receptor-mediated dimers in the crystal prompted us to study dimerization in solution and to contemplate its role in signal transduction on the membrane.

Assembly of the FSH-FSHR_{HB} complex

Human FSH and a truncated ectodomain of human FSHR were co-secreted as a stable complex from insect cells using the baculovirus expression system. The α - and β -subunits of the heterodimeric FSH were joined into a single chain with a Gly/Ser-rich linker from the carboxy terminus of the β -subunit to the amino terminus of the α -subunit; similar FSH and CG constructs, but with different linkers, have been shown to retain a wild-type level of function^{11–13}. All glycoprotein hormone receptors comprise a series of LRRs flanked at each end by cysteine clusters. Refolding experiments of recombinant LHR indicate that the N-terminal cysteine cluster is required for proper folding (data not shown). A truncated FSHR consisting of the N-terminal cysteine cluster and the LRRs forms a stable complex with FSH, indicating that the C-terminal cysteine-rich region of the ectodomain is not required for hormone binding. We therefore define the N-terminal cysteine cluster together with the LRRs as the hormone-binding domain of the receptor (FSHR_{HB}).

Structure determination

Co-secreted FSH and FSHR_{HB} form a 1:1 complex as demonstrated by its retention time upon gel filtration chromatography (data not shown). The fully glycosylated FSH-FSHR_{HB} complex could be crystallized in several different forms, the best of which only diffracted to 9 Å spacings. Partial deglycosylation using a combination of endoglycosidases F2 and F3 yielded a new crystal form, which diffracted to 2.9 Å spacings. The crystals are of space group C2, with two complexes in the asymmetric unit. The complex structure was determined by molecular replacement. The positions of the two hormone molecules were located using the known structure of human FSH⁷ as the search model. Partial phases obtained from the molecular replacement solution were improved by two-fold molecular averaging to yield electron density for the

bound receptor. A complete model of the complex was developed iteratively and refined ultimately at 2.9 Å resolution to an *R*-value of 21.9% (*R*_{free} = 25.9%).

Receptor structure consisting of irregular repeats

FSHR_{HB} is formed from its LRR winding into the shape of a slightly curved tube (Fig. 1a, b; see also Supplementary Fig. S1). As predicted from sequence considerations¹⁴ and as found in other LRR structures¹⁵, its concave inner surface is an untwisted, non-inclined β-sheet and its tubular hydrophobic core is filled by LRR namesake leucines and other large hydrophobic residues. The repeats themselves are highly irregular in length and conformation, and particularly so in the outer surface returns. Regular structure in the convex outer surface is limited to seven short β-strands in three sheets (Fig. 1c). Unexpectedly, the N-terminal flanking sequence (residues 18–46) contains the first of the LRR repeats. This segment is further integrated into the structure by an antiparallel β-strand added before the inner β-sheet and by two disulphide bridges (Cys 18–Cys 25 and Cys 23–Cys 32) (Fig. 1a; see also Supplementary Fig. S2).

Unlike other known LRR structures, the curvature of FSHR_{HB} is steeply graded; the C-terminal portion (repeats 7–10) has the signature horseshoe-like curvature of LRR proteins whereas the N-terminal portion of the inner β-sheet (repeats 1–7) is nearly flat. Apart from repeat 1, the overall fold of FSHR_{HB} is likely to be conserved in LHR and TSHR, and across mammalian species, because Leu/Ile/Phe residues in the hydrophobic core of FSHR_{HB} are conserved at equivalent positions in the sequences from LHR and TSHR (Fig. 2c) and almost all repeat irregularities are also conserved. We are describing FSHR_{HB} structural details and comparisons to other LRR proteins elsewhere.

A hand-clasp binding mode

The structure of the FSH–FSHR_{HB} complex shows FSH bound into the concave face of the curved receptor domain in a manner that resembles a hand clasp (Figs 1a, b and 2a), roughly as in our prediction¹⁴. The long axis of FSH is perpendicular to the FSHR_{HB} tube and aligned with receptor β-strands. All ten parallel β-strands and the loops just C-terminal to the β-sheet are in contact with FSH (Fig. 3a; see also Supplementary Fig. S1). The receptor wraps around the middle section of the hormone molecule, interacting with C-terminal segments of both FSH-α and FSH-β as well as the αL2 and βL2 loops. The hormone orientation is such that loops βL1 and βL3 extend out from the C-terminal tips of the receptor

inner sheet, whereas loops αL1 and αL3 protrude away from the base.

The footprints of hormone and receptor onto their partner surfaces (Fig. 2a; see also Supplementary Fig. S1) demonstrate the substantial involvement of both FSH subunits in the interaction with FSHR_{HB}. A total of 2,600 Å² of solvent-accessible surface area from the isolated components is buried into the interface between FSH and FSHR_{HB}. Most of this interface involves the packing of two rather flat surfaces: one formed by the seven N-terminal parallel β-strands of FSHR_{HB} is matched with one from the C-terminal segment of FSH-α and the C-terminal ‘seat-belt’ segment of FSH-β. The rest comes from the αL2 loop with its helical segment clamping over the edge of FSHR_{HB} at the β-strand tips of LRRs 2–6, and from the βL2 loop fitting against the curved C-terminal repeats 8–10. A significant consequence of this geometry is that a swathe of the most deeply buried interfacial receptor residues (Fig. 2c; see also Supplementary Fig. S3) lies across the C-terminal end of the parallel β-sheet (Fig. 3a). Half of these FSHR_{HB} residues (L55, N79, K104, N129, Q152, D153 and K179) are buried by contacts with both FSH-α and FSH-β (Fig. 2a, magenta).

The interface between hormone and receptor is bare of carbohydrate groups. Sufficient density was found to model carbohydrate residues (1–3) at all four potential glycosylation sites on the hormone (FSH-α N52 and N78; FSH-β N7 and N24) and at the one verified glycosylation site on our receptor construct (N191)¹⁶. All of these sites are remote from the interface (Fig. 1a, b). Even the closest carbohydrate moiety, at FSH-α N52, faces away from the receptor interface. Glycosylation at FSH-α N52 has been implicated in signalling but not in binding². Natural mutations located at or close to the conserved glycosylation site of FSHR (N191I, A189V) have been shown to cause inactivation of the receptor², possibly by affecting the proper folding of FSHR¹⁶.

Charge complementarity in a universal binding interface

Several lines of evidence suggest that the mode of binding seen here between FSH and FSHR_{HB} probably also pertains to other mammalian glycoprotein hormones and their receptors. First, highly similar structures for human FSH⁷ and CG^{5,6}, and highly conserved sequences for LH and TSH, including identical α-chains within a species, point to a common ligand structure in the complex. Second, for reasons noted above, we expect the structure of the hormone-binding domain to be essentially the same in all glycoprotein hormone receptors. Third, previous studies on the different systems^{1,8,9} implicate the very structural elements found in our FSH–

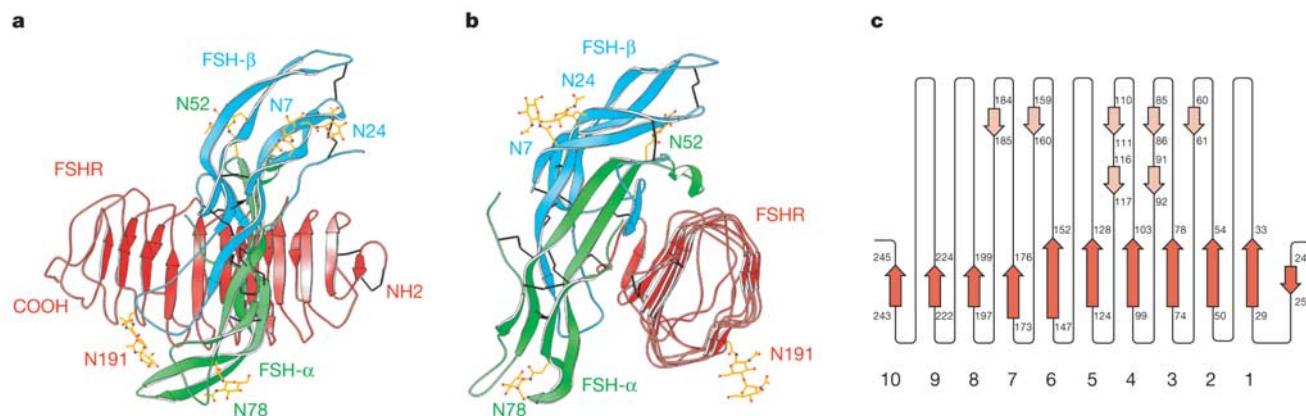


Figure 1 Crystal structure of human FSH bound to FSHR_{HB}. **a, b**, Ribbon diagram of the complex structure shown in two views related by a 90° rotation about the vertical axis. FSH α-chains and β-chains are in green and cyan, respectively. FSHR_{HB} is in red. The observed N-linked carbohydrates at N52 and N78 of FSH-α, N7 and N24 of FSH-β, and

N191 of FSHR_{HB} are in yellow. Disulphide bonds are in black. **c**, Schematic diagram of the topology of FSHR_{HB} structure. β-Strands are shown as arrows. Red represents strands located at the concave face of FSHR_{HB}; pink represents strands on the convex face.

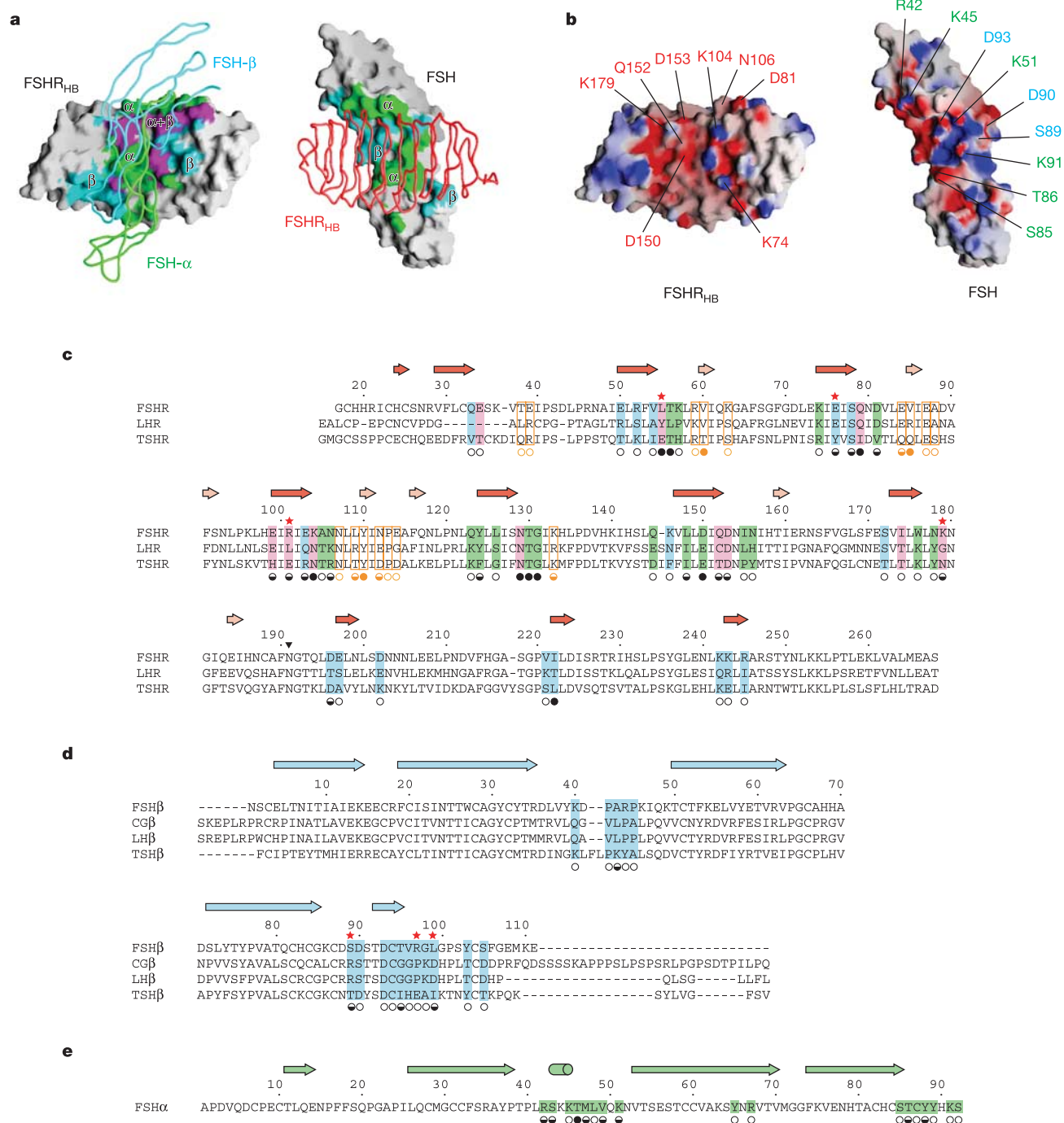


Figure 2 Recognition of FSH by FSHR_{HB}. **a**, The left panel shows the molecular surface of FSHR_{HB} with the imprint of bound FSH (C α trace in green for α - and cyan for β -chains). Coloured patches represent FSHR_{HB} residues that are buried at the receptor–hormone interface by FSH- α alone (green), by FSH- β alone (cyan), or both FSH- α and FSH- β (magenta). The right panel shows the molecular surface of FSH with the imprint of FSHR_{HB} (C α trace in red). Green denotes residues of FSH- α buried by FSHR_{HB}; cyan represents buried FSH- β residues. **b**, Electrostatic potential surface of FSHR_{HB} (left) and FSH (right). Residues of FSHR_{HB} are marked red; residues of the FSH- α and FSH- β chains are in green and cyan, respectively. **c–e**, Sequence alignment. Secondary structure assignments are shown as arrows for β -strands and cylinders for α -helices. For each residue buried at the receptor–ligand interface or the receptor dimer interface, the fractional solvent accessibility is indicated by an open circle if it is greater than 0.4, a half-

filled circle if it is 0.1–0.4, and a filled circle if it is less than 0.1. Residues implicated in binding specificity are marked by asterisks. **c**, Human FSHR, LHR and TSHR sequences in the region of the hormone-binding domain. β -Strands located at the concave face of FSHR are coloured red, whereas strands on the convex face are in pink. FSHR_{HB} residues buried at the receptor–ligand interface by FSH- α alone (green), FSH- β alone (cyan), or both FSH- α and FSH- β (magenta) are highlighted. FSHR_{HB} residues buried at the receptor dimer interface are boxed in orange. N-linked glycosylation is marked by a black triangle. **d**, Human FSH, CG, LH and TSH β -chain sequences. FSH- β residues buried at the receptor–ligand interface are highlighted in cyan. **e**, The common human α -chain sequence. FSH- α residues buried at the receptor–ligand interface are highlighted in green.

FSHR_{HB} complex to be involved in the interaction: namely the LRRs of the receptor and the C-terminal segments of the hormone subunits. Last, residues in the receptor homologues at positions where FSHR_{HB} contacts FSH- α are consistent with a common mode of binding (Fig. 2c). Many of these residues are conserved, and several are charged groups interacting with complementary groups on the hormone.

The interface between FSH and FSHR_{HB} has an exceptionally high buried-charge density in comparison to other protein complexes (1.13 charges per nm²; Q.F. & W.A.H., unpublished data). The electrostatic potential surfaces of the hormone and receptor in the interface are predominately positive and negative, respectively, although less uniformly than in the putative CG-LHR complex¹⁴. The electropositive and electronegative patches in the interface reflect direct interactions between residues with complementary charges (Fig. 2b), many of which are conserved in the homologues. Two invariant acidic residues (D150 and D153 in FSHR) make salt bridges with basic residues from FSH- α (K91 and K51). In addition, D81 from FSHR_{HB} (also Asp in LHR) forms charge-mediated interactions with R42 and K45 of the α L2 loop of FSH. These findings are consistent with previous studies. Charge reversal mutagenesis of E132 and D135 individually to Lys in LHR (residues corresponding to D150 and D153 in FSHR) abolished hCG binding¹⁷; mutation of FSH- α residues K51 (ref. 18) or K91 (ref. 19) to Ala caused the loss of receptor-binding activity; and inhibition of hormone binding using synthetic peptides indicated that the α L2 regions of both human CG and TSH are involved in receptor binding^{20,21}.

Specificity jointly mediated by hormone α - and β -subunits

Having common α -chains within a species, glycoprotein hormones

must be distinguished by their β -chains. Important discriminating determinants of receptor binding have been localized to β -chain C-terminal segments^{22–25} that correspond to a unique seat-belt feature of the hormone structures^{5–7}. The C-terminal end of the β -subunit embraces the α -subunit and fastens into place via a disulphide tether back to the body of the β -subunit^{5–7}. The seat-belt segment of FSH- β (89–105, Fig. 2d) is indeed at the interface, but it is sandwiched between the α L2 loop and C-terminal segment of FSH- α (Fig. 3a; see also Supplementary Fig. S4). Specificity pockets for residues of FSHR_{HB} form through coordination between discriminating β -chain seat-belt residues and common α -chain residues. The only other part of FSH- β at the interface is the tip of loop β L2 (40–45).

We have analysed interfacial contacts for potential determinants of specificity (Supplementary Table S1 and Fig. S5). Five substantially buried (<40% exposed) FSHR_{HB} residues (L55, E76, R101, K179 and I222; Fig. 2c) both vary among human receptors and make contact with FSH residues that vary among the counterpart hormones. Sequence variations indicate that hormone seat-belt interactions with receptor residues at positions 55 and 179 determine the specificity between FSH and TSH versus LH/CG, whereas residues at positions 76 and 101 distinguish between FSH and TSH. The interaction between position I222 and β L2 does not appear to be strongly discriminating. Other variable residues that are substantially buried in FSHR_{HB} (Q79, E103, K104, Q152 and D196; Fig. 2c) are covered in part by FSH- β but are not in direct contact with it; however, they could be selective in homologous complexes. Although collective interfaces must contribute to specificity, three sites appear to predominate. These focus on receptor residues corresponding to L55, K179 and the combination of E76 and R101 in FSHR.

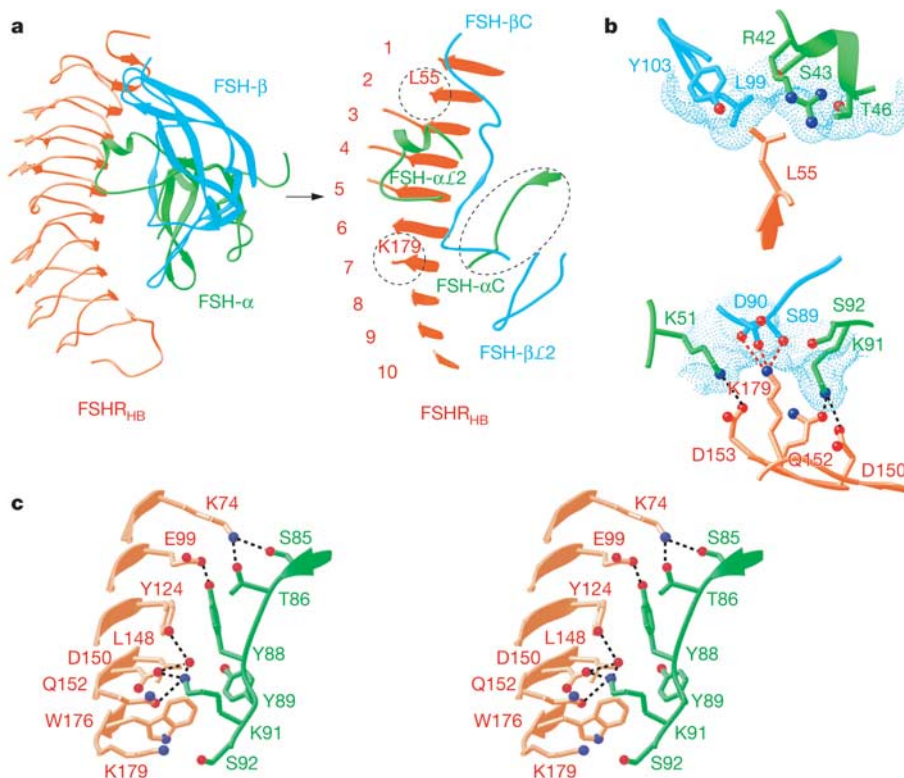


Figure 3 Interactions at the receptor–ligand interface. **a**, Ribbon diagram showing the top view of the FSH–FSHR_{HB} complex. The view in the right panel is tilted to highlight the regions of FSHR_{HB} (red), and FSH- α (green) and FSH- β (cyan) chains that are involved in direct contacts at the receptor–ligand interface. Dashed circles mark the locations of L55

and K179 in the FSHR_{HB} structure. **b**, Detailed views of the interactions at the specificity pockets for L55 and K179 of FSHR_{HB}. The dotted molecular surface of FSH is shown in cyan in each panel. **c**, Close-up stereo view of the interactions between the C-terminal region of FSH- α and FSHR_{HB}.

The molecular surface of FSH provides a shallow pocket for the side chain of L55, which forms hydrophobic interactions with L99 and Y103 of FSH- β and with the aliphatic part of R42 from FSH- α (Fig. 3b). Substitution of L55 in FSHR with the Tyr residue found in LHR resulted in marked loss of function and hormone binding for the mutant receptor²⁶, and it is clear from the structure that steric hindrance would preclude accommodation of the larger Tyr residue in the shallow pocket for L55. The residues in LH or CG (D105 and T109) that correspond to L99 and Y103 in FSH- β are smaller and could potentially form polar interactions or hydrogen bonds with the Tyr residue found in LHR. Conversely, these polar groups in LH/CG would not be complementary to the hydrophobic L55 of FSHR, and indeed FSH binding to FSHR is impaired when residues 95–100 of FSH- β (TVRGLG) are replaced with residues 101–106 of LH- β (GGPKDH)²⁴.

A channel on the molecular surface of FSH hosts the side chain of FSHR K179 (Fig. 3b). The sides of this channel come from K51 and K91 of FSH- α , which make the universal salt bridges discussed above, and its top is capped by S89 and D90 of FSH- β , which form three hydrogen bonds with K179 of FSHR. S89 is substituted with R95 in the β -subunits of CG/LH (Fig. 2d). Adverse interactions of this Arg residue with K179 would be expected to destabilize a potential complex between CG/LH and FSHR. Thus, the structure affords an explanation for extensive mutagenesis studies showing that K179 of FSHR functions as a negative determinant to prevent the binding of CG/LH to FSHR^{26,27}; for example, mutation of K179 to the Gly residue found in LHR gave the mutant FSHR a gain of sensitivity towards hCG²⁶. Moreover, FSH acquires LHR binding activity when residues 88–91 of FSH- β (DSDS) are replaced with their LH- β counterparts (94–97; RRST)²⁴.

A common feature of the specificity pockets for L55 and K179 is the involvement of residues from both the α - and β -chains of FSH, combining stereochemical orientation with specific interactions to mediate selectivity. The third site of selectivity involves the α -chain less directly. Structurally adjacent FSHR_{HB} residues E76 and R101 (Supplementary Fig. S5) interact primarily with R97 and V96 (respectively) from FSH- β , although FSH- α segments contribute to solvent exclusion. Charged side chains do not make hydrogen-bonding contacts here, but repulsion expected from FSHR residues E76 and R101 interacting with the TSH- β Glu and His counterparts of R97 and V96 might explain FSH versus TSH selectivity. Although TSH chimaeras substituted with the FSH- β seat-belt segments failed

to elicit follitropic activity, their reduced thyrotropic responses are compatible with our predictions²⁵. The lengthened β L2 loop of TSH may also participate in selectivity.

Receptor-induced conformational changes in hormone

Although the free and receptor-bound structures of FSH are generally similar, the hormone conformation is substantially changed and appreciably rigidified in the complex (Fig. 4). Free FSH is evidently quite flexible (C α root-mean-squared deviation (r.m.s.d.) of 1.85 Å between copies⁷), and this is especially true at termini and for the α L1, α L3 and β L2 loops at one end. The C terminus of FSH- α and these basal loops are the elements subject to receptor-induced conformational changes, which move them into a fixed conformation (C α r.m.s.d. of 0.71 Å between copies) distinct from any in the free state (C α r.m.s.d. of 1.5–2.9 Å overall; 2.6–6.3 Å for the C terminus of FSH- α and the loops α 13–26, α 66–78, α 88–92 and β 36–47). The conformation of free CG^{5,6} is also distinct from bound FSH.

The most marked change occurs at the C terminus of FSH- α where the terminal segment (88–92) is rotated almost 180° to place its end more than 20 Å away from the free-state position. This segment is buried into the receptor interface where it is fully ordered, whereas the last two residues are disordered in free FSH⁷, as are the last three in hCG^{5,6}. Extensive interactions with the receptor include five direct hydrogen bonds, water-mediated contacts, and hydrophobic interactions (Fig. 3c). Many of these contacts involve residues that are conserved among glycoprotein hormone receptors, consistent with a universal mode of binding.

A key feature of involvement of the FSH- α C terminus in the interface is an aromatic ring-stacking interaction between Y88 of FSH- α (Y127 in human LHR and F130 in TSHR) and Y124 of the receptor. Y88 also forms a hydrogen bond with E99 of FSHR. Altogether, 106 Å² of solvent-accessible surface area from Y88 on FSH- α is buried into the interface. Y88 is the first residue to diverge significantly from the free-state conformation, and its importance in the receptor interaction is demonstrated by experiments where deletion of the terminal residues 88–92 of hCG- α essentially eliminates receptor binding, whereas deletion of just the last four (89–92) retains binding albeit with reduced potency²⁸. Alanine mutagenesis at the C terminus of FSH- α showed that six contiguous residues including Y88 contribute to the FSH–FSHR interaction¹⁹.

The basal loops are locked into their receptor-bound state,

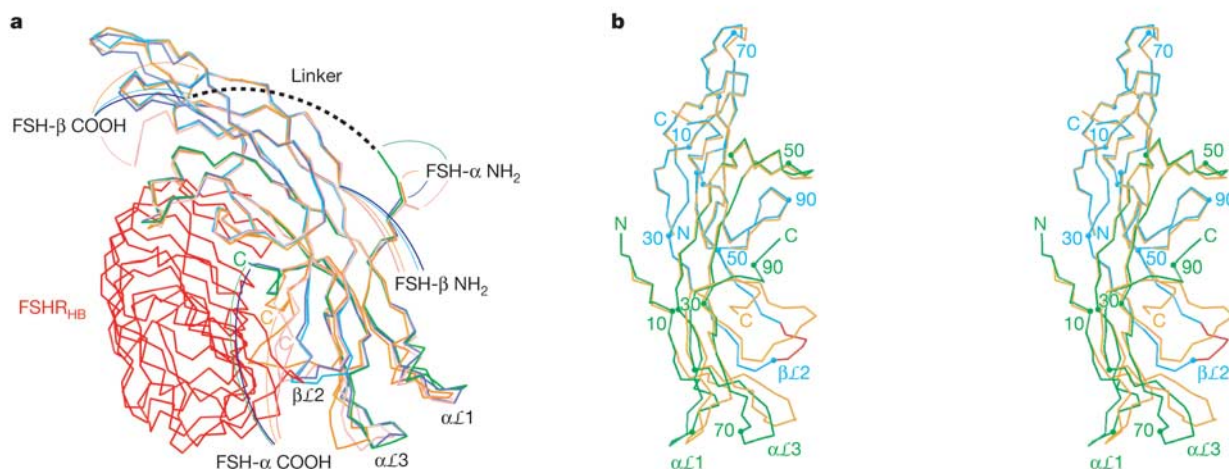


Figure 4 Different conformations in free and receptor-bound FSH. **a**, C α trace superposition of four independent copies of FSH. One protomer of FSH is bound to the FSHR_{HB} shown in red; its α -chain is in green and β -chain in cyan. The second copy of receptor-bound FSH is in blue. The two protomers of free FSH are in orange and pink.

b, Stereo diagram of the superposition of one receptor-bound FSH (α -chain in green; β -chain cyan) with one free hormone (orange). The view is rotated 90° about the vertical axis from **a**. Every 20th residue of the receptor-bound FSH is marked. Red highlights the segment of the β L2 loop that is involved in receptor binding.

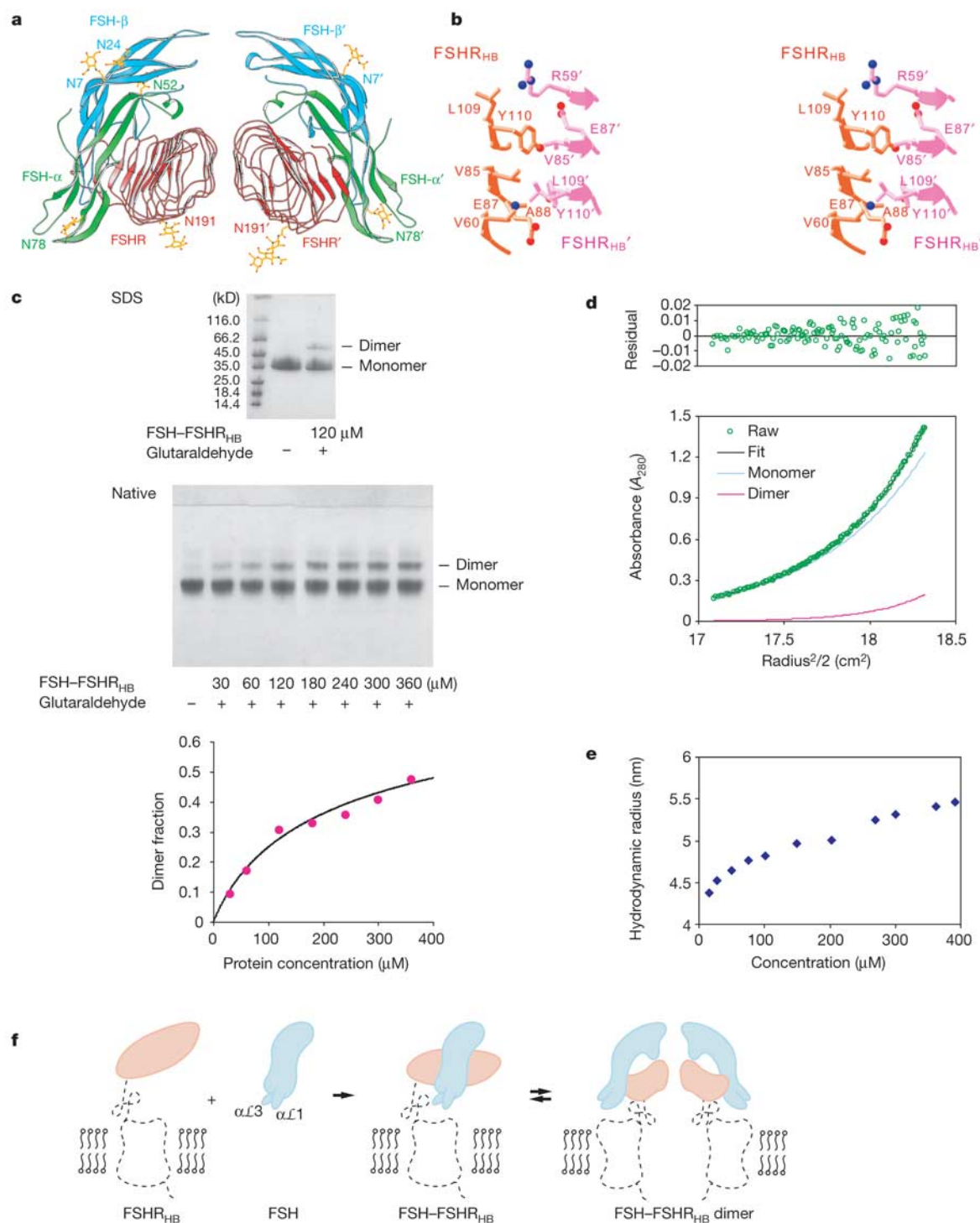


Figure 5 Dimer of the FSH-FSHR_{HB} complex. **a**, Ribbon diagram of a non-crystallographic dimer of the FSH-FSHR_{HB} complex (FSH-FSHR_{HB} and FSH'-FSHR_{HB}'). The receptor is in red, and the α - and β -chains of the hormone are in green and cyan, respectively. **b**, Stereo view of the direct contacts (distances $< 4 \text{ \AA}$) observed at the dimer interface. **c**, Crosslinking of the FSH-FSHR_{HB} complex by glutaraldehyde. The SDS gel contains the FSH-FSHR_{HB} control (lane 2) and 120 μM FSH-FSHR_{HB} complex reacted with 120 μM glutaraldehyde (lane 3). The native gel contains the FSH-FSHR_{HB} control (lane 1), and FSH-FSHR_{HB} complex reacted with glutaraldehyde at a 1:1 molar ratio and at concentrations of 30, 60, 120, 180, 240, 300 and 360 μM (lanes 2–8, respectively). Densities from the dimer bands on native gels as a function of protein concentration were fitted to the dimerization equation with a single K_d ($453 \pm 140 \mu\text{M}$) for six experiments, with two repeated trials for each molar ratio of FSH-FSHR_{HB} to glutaraldehyde (3:1, 3:2, 1:1). The average dimer fraction at each protein concentration is plotted together with the

fitted curve. **d**, Analytical ultracentrifugation analysis of dimer formation by the FSH-FSHR_{HB} complex. Shown are equilibrium sedimentation data (green open circle) taken at 12,000 r.p.m. with a loading concentration of 1.05 mg ml^{-1} , and the best-fit curve (black solid line through the data) for a monomer-dimer reversible equilibrium model. Under the fit, the theoretical sum is deconvoluted into contributions from the monomer (blue line) and dimer (red line) species, respectively. Residuals in absorbance units at 280 nm are shown at the top of the plot. **e**, Apparent hydrodynamic radii of FSH-FSHR_{HB} measured by dynamic light scattering at concentrations of 15, 27, 50, 75, 100, 150, 200, 270, 300, 360 and 390 μM . **f**, Model for FSHR activation by FSH. Projections of the FSH-FSHR_{HB} complex and bovine rhodopsin structures are used to represent the hormone (cyan), the hormone-binding domain (orange) and the linker and transmembrane domain (black dotted trace) of the receptor. The $\alpha\text{L}1$ and $\alpha\text{L}3$ loops of FSH are proposed to contact the transmembrane domain directly.

presumably one favourable for GPCR activation, through interactions that begin with FSH β L2 contacts in the complex. Multiple free-state conformations of β L2 would conflict with the receptor in the complex and with the receptor-bound conformation of the FSH- α C terminus. Instead, β L2 adopts a new and fixed conformation by interacting positively with the three C-terminal LRRs of FSHR_{HB} (Supplementary Table S1 and Fig. S4). Then, it seems that contacts of β L2 with its α L3 neighbour move that loop elastically by up to 7.0 Å. In turn, contacts of α L3 with α L1 press the elastic deformation into α L1 to define the final bound-state conformation.

Dimerization of the FSH-FSHR_{HB} complex

The two independent complexes of FSH-FSHR_{HB} in the crystal interact to constitute a dimer. A quasi-two-fold axis of rotational symmetry (rotation $\chi = 175.2^\circ$, screw translation $t_\chi = -1.2$ Å) relates the two protomers, passing between receptor molecules approximately perpendicular to the tube axes, such that the tips of α L1 and α L3 from the hormone are coplanar (Fig. 5a). The dimer interface involves the three-stranded β -sheet (LRRs 2–4) on the outer surface of FSHR_{HB} (Fig. 5b). Y110 (Fig. 2c) contributes prominently in a largely hydrophobic interaction, mediating 21 out of 24 direct carbon-carbon contacts. Because Y110 is completely conserved (Fig. 2c), the homologous receptors may form dimers in a similar manner. The solvent-accessible surface area buried at the dimeric interface is relatively small (940 Å² total from both non-hydrated protomers), which suggests a weak association.

Dimers of the FSH-FSHR_{HB} complex were detected in solution by chemical crosslinking, analytical centrifugation and light scattering. Crosslinking of FSH-FSHR_{HB} with equal or less than equal molar ratios of glutaraldehyde produced dimer bands on SDS and native gels in a concentration-dependent manner. Quantification of dimer formation as a function of protein concentration could be fitted with a K_d of 453 ± 140 μ M (Fig. 5c). Sedimentation equilibrium experiments indicated that the FSH-FSHR_{HB} complex exists in a monomer-dimer equilibrium (Fig. 5d) with a K_d of 388 ± 120 μ M. Dynamic light-scattering measurements for FSH-FSHR_{HB} showed that the apparent hydrodynamic radius increased 25% over the concentration range of 15–390 μ M (Fig. 5e), well beyond the concentration dependence found for non-associating proteins.

Implications for receptor activation

We expect the picture of ligand binding provided by the FSH-FSHR_{HB} complex to apply to all glycoprotein hormone systems, as discussed above. Moreover, glycoprotein hormone receptors have 7TM portions that are very similar, both to one another (~70% sequence identity) and to other rhodopsin-family GPCRs, and when activated they can directly stimulate the same G_s protein used by other GPCRs. Thus, we anticipate a mechanism for receptor activation that is essentially the same for FSHR, LHR and TSHR, and one that has elements in common with other GPCRs²⁹. Now that we have structures in hand for an ectodomain complex with an activating ligand and for rhodopsin³⁰ as an inactive-state model of the 7TM domain, we are in a position to contemplate receptor activation in this system. To do this, we must first understand how the FSH-FSHR_{HB} complex is oriented relative to the 7TM portion. Mapping experiments show that hormone epitopes on both the α -tip (α L1, α L3) and β -tip (β L1, β L3) are antibody accessible in hCG-porcine LHR (pLHR) ectodomain complexes but accessibility to the α -tip is occluded in complexes with full-length pLHR³¹. Thus, we believe that the complex is oriented as in Figs 1, 4 and 5 relative to a cell surface below it: that is, with α -tips membrane-proximal and the receptor dyad approximately normal to the surface.

At least three models have been suggested for the activation of glycoprotein hormone receptors⁸. One invokes hormone-induced

conformational changes in the ectodomain that directly activate the 7TM domain³²; a second features ectodomain constraints on activity that are released by hormone binding or by certain mutations³³; and a third proposes that it is the hormone when bound to the ectodomain that directly activates the 7TM domain^{14,31,32}. Elements of each might be involved. Although aspects of hormone binding and receptor activation are separable, these models imply a necessary thermodynamic linkage.

We do not yet know the structure of free FSHR_{HB}, but based on other LRR structures¹⁵ it seems unlikely that any substantial changes in this domain accompany hormone binding (model 1). Certainly nothing of the kind found in the dimeric ectodomains of metabotropic glutamate receptors³⁴ is feasible here. Moreover, the finding that LHR/FSHR chimaeras with swapped extracellular domains respond fully in accord with hormone-binding selectivity³² implies that membrane-proximal surfaces of the hormone-binding domain, which are quite variable, are not involved in activation.

The model of hormone-released inhibition (model 2) derives from observations of ectodomain mutations that confer hormone-independent, constitutive activity to the receptor³³. These mutations, at receptor counterparts of S273 just beyond the end of FSHR_{HB}, are in a conserved segment in the linker connecting the LRRs to the 7TM domain. Apart from conservation at each end and at six cysteine positions, which are thought to create a condensed structure by disulphide bridging, these linkers are highly variable among receptors in sequence and length (87–137 residues). Whether S273 mutations truly release inhibition or activate positively is unclear because receptors lacking ectodomains are inactive⁸ or not fully active³³. Semantics aside, involvement of the linker region in activation is apparent both from these mutations and from the requirement for tyrosine sulphation at a site towards the membrane end of the linker³⁵. The α L1/ α L3 tips in the complex protrude towards a space where they might interact with the linker structure in normal, hormone-directed signalling.

The FSH-FSHR_{HB} structure is fully compatible with direct interactions between the α -tips of the hormone in its receptor-bound conformation and the 7TM domain and linker elements (model 3). The β L2 loop would seem to be hidden from direct contact, but its integrity would be essential to generate the bound conformation of α L1 and α L3 (Fig. 4). Such activation might be mediated either by contacts with 7TM loops or, plausibly, by insinuation into the 7TM barrel after displacement of the second extracellular loop (EL2)³⁰. Several experiments are consistent with this model, including the chimaeric exchanges³², antibody occlusion³¹, EL2-peptide inhibition of signalling and α -tip antibody access³¹, and signal-enhancing α L1 mutations³⁶. Moreover, the remoteness of the carbohydrate at FSH- α N52 from conceivable activation sites is consistent with results from disulphide-stabilized hCG showing no effect on activity as a result of mutational deglycosylation at this site³⁷. On the other hand, the structure is incompatible with experiments implicating the C terminus of the α -subunit in receptor activation³⁸.

Our observation of dimerization by the FSH-FSHR_{HB} complex suggests that such dimerization may also have a role in glycoprotein hormone receptor function (Fig. 5f). Indeed, dimerization is increasingly found to be important for signal transduction by GPCRs³⁹. Although the dimeric interaction seen for soluble FSH-FSHR_{HB} is quite weak ($K_d \sim 400$ μ M), this affinity is expected to be enhanced substantially when the receptors are anchored in the membrane⁴⁰. Dimerization of hormone-receptor complexes is consistent with the co-expression of mutant receptor defective in hormone binding with others defective in signal transduction, whereby ligand-induced cAMP production is reconstituted^{41,42}, and also with direct measures of self-association through co-immunoprecipitation of differentially tagged receptors⁴³ or by fluorescence-resonance energy transfer experiments⁴⁴. Puzzles do remain, however, as receptor C termini in the dimeric complex are

separated by a distance (74.3 Å) that, depending on linker flexibility, may be too great to accommodate contacts between associated 7TM domains (Fig. 5f).

Methods

Protein expression and purification

We made a pFastBac Dual plasmid (Invitrogen) to encode both human FSH and human FSHR_{HB} under control of separate promoters for introduction into a baculovirus vector. The hormone construct comprises the signal peptide and entire mature sequence of FSH-β (residues 1–111) followed by a 15-residue linker (GGGSGGGSGGGSGGG) and the mature sequence of FSH-α (residues 1–92). The receptor construct comprises the native signal sequence followed by the FSHR_{HB} ectodomain (residues 1–268) and a Flag tag to facilitate purification. Hi5 cells infected with this recombinant baculovirus secreted the FSH–FSHR_{HB} complex, which we purified from the cell supernatant by affinity chromatography (anti-Flag antibody M2) and size exclusion chromatography (Superdex200). The yield of purified FSH–FSHR_{HB} was about 0.1 mg per litre of insect cell culture. We partially deglycosylated the complex using endoglycosidases F2 and F3, and purified the product by ion exchange chromatography (MonoQ).

Crystallization and data collection

Fully glycosylated FSH–FSHR_{HB} were crystallized from tert-butanol into large polyhedra. These crystals are in space group R32 (*a* = 227.2 Å, *α* = 80.3°) with an estimated five to ten complexes per asymmetric unit, but they diffract only to 9 Å spacings (NSLS beamline X4A). Partially deglycosylated FSH–FSHR_{HB} crystallized from PEG3350 and dilute Li₂SO₄ as clusters of thin plates (5–10 μm thick). These crystals belong to space group C2 (*a* = 121.3 Å, *b* = 66.9 Å, *c* = 148.6 Å and *β* = 99.1°) and have two FSH–FSHR_{HB} complexes per asymmetric unit. Diffraction extended to 2.9 Å spacings at APS beamline 31-ID of SGX-CAT, where we measured complete data that were reduced using Denzo and Scalepack⁴⁵ (Table 1).

Structure determination and refinement

We used molecular replacement and molecular averaging to solve the structure. A two-fold non-crystallographic symmetry (NCS) axis was identified from the self-rotation function, and the known structure of FSH⁷ was used as the search model in Amore⁴⁶ to locate both FSH molecules in the asymmetric unit. After rigid-body refinement against all data (25–2.9 Å), phases from this model were improved by two-fold NCS averaging in DM⁴⁷. The resulting electron-density map revealed a β-sheet in the receptor region as predicted for the LRR motif of FSHR_{HB}. We developed a complete atomic model of the complex through a succession of manual fittings and iterative averaging. Optimal NCS restraints were determined from an *R*_{free} analysis (Q.F. & W.A.H., unpublished data). We used O⁴⁸ for model building, CNS⁴⁹ for initial refinement and REFMAC⁴⁷ in the final stages. Refinement statistics are given in Table 1.

The final model contains two FSH molecules (residues α3–92 and α6′–92′; β3–107 and β3′–107′), two FSHR_{HB} molecules (residues 18–259 and 18′–250′), one mannose and ten *N*-acetylglucosamine residues, 50 water molecules, and three sulphate ions. Continuous density was observed for carbohydrate moieties at all N-linked glycosylation sites on one FSH–FSHR_{HB} complex (αN52, αN78, βN7 and βN24 of FSH; N191 of FSHR_{HB}) and three sites on the other (αN78′ and βN24′ of FSH′; N191′ of FSHR_{HB}′). There was no density for residues of the Flag-tagged peptide and the FSH-β–FSH-α linker, presumably due to disordering. A Ramachandran analysis (<http://kinemage.biochem.duke.edu>) places 92.1% of all residues in favoured regions and 0.9% in outlier regions.

Chemical crosslinking

We carried out reactions of the FSH–FSHR_{HB} complex at concentrations up to 360 μM

with fixed molar ratios of glutaraldehyde that were sufficiently low to prevent adventitious crosslinking at higher protein concentrations (30 lysines per complex). A concentration-dependent dimer band on native gels was quantified by densitometry. The dimer-band density (*D*) from this crosslinked subpopulation is proportional to the equilibrium dimer fraction, *D* = *Sy*, where *S* is a scaling factor, and the dimer fraction (*y*) depends on the total protein concentration (*C*_T) and *K*_d according to

y = (sqrt(1 + 8C_T/K_d) - 1) / (sqrt(1 + 8C_T/K_d) + 1)

A least-squares fitting of the *D*(*C*_T) data gave estimates of *S* and *K*_d.

Analytical ultracentrifugation

We carried out sedimentation equilibrium experiments using absorption optics in a Beckman XL-I analytical centrifuge. Purified fully glycosylated FSH–FSHR_{HB} was analysed at three concentrations (0.33, 0.66 and 1.05 mg ml^{−1}) and three rotor speeds (7,000, 9,000 and 12,000 r.p.m.). Data were analysed with WinNONLIN⁵⁰, floating the monomer mass to determine the extent of glycosylation. The FSH–FSHR_{HB} mass was found to be 78 kDa, implying 32% carbohydrate mass.

Dynamic light scattering

We measured hydrodynamic radii of FSH–FSHR_{HB} using a DynaPro molecular sizing instrument (Protein Solutions).

See supplementary information for details of the Methods.

Received 27 August; accepted 22 November 2004; doi:10.1038/nature03206.

1. Dias, J. A. *et al.* Molecular, structural, and cellular biology of follitropin and follitropin receptor. *Vitam. Horm.* **64**, 249–322 (2002).

2. Simoni, M., Gromoll, J. & Nieschlag, E. The follicle-stimulating hormone receptor: biochemistry, molecular biology, physiology, and pathophysiology. *Endocr. Rev.* **18**, 739–773 (1997).

3. Themmen, A. P. N. & Huhtaniemi, I. T. Mutations of gonadotropins and gonadotropin receptors: elucidating the physiology and pathophysiology of pituitary-gonadal function. *Endocr. Rev.* **21**, 551–583 (2000).

4. Pierce, J. G. & Parsons, T. F. Glycoprotein hormones: structure and function. *Annu. Rev. Biochem.* **50**, 465–495 (1981).

5. Wu, H., Lustbader, J. W., Liu, Y., Canfield, R. E. & Hendrickson, W. A. Structure of human chorionic gonadotropin at 2.6 Å resolution from MAD analysis of the selenomethionyl protein. *Structure* **2**, 545–558 (1994).

6. Laphorn, A. J. *et al.* Crystal structure of human chorionic gonadotropin. *Nature* **369**, 455–461 (1994).

7. Fox, K. M., Dias, J. A. & Van Roey, P. Three-dimensional structure of human follicle-stimulating hormone. *Mol. Endocrinol.* **15**, 378–389 (2001).

8. Ascoli, M., Fanelli, F. & Segaloff, D. L. The lutropin/choriogonadotropin receptor, a 2002 perspective. *Endocr. Rev.* **23**, 141–174 (2002).

9. Szkudlinski, M. W., Fremont, V., Ronin, C. & Weintraub, B. D. Thyroid-stimulating hormone and thyroid-stimulating hormone receptor structure-function relationships. *Physiol. Rev.* **82**, 473–502 (2002).

10. Hsu, S. Y. T. New insights into the evolution of the relaxin-LGR signaling system. *Trends Endocrinol. Metab.* **14**, 303–309 (2003).

11. Sugahara, T. *et al.* Biosynthesis of a biologically active single peptide chain containing the human common α and chorionic gonadotropin β subunits in tandem. *Proc. Natl Acad. Sci. USA* **92**, 2041–2045 (1995).

12. Narayan, P., Wu, C. & Puett, D. Functional expression of yoked human chorionic gonadotropin in baculovirus-infected insect cells. *Mol. Endocrinol.* **9**, 1720–1726 (1995).

13. Schmidt, A., MacColl, R., Lindau-Shepard, B., Buckler, D. R. & Dias, J. A. Hormone-induced conformational change of the purified soluble hormone binding domain of follitropin receptor complexed with single chain follitropin. *J. Biol. Chem.* **276**, 23373–23381 (2001).

14. Jiang, X. *et al.* Structural predictions for the ligand-binding region of glycoprotein hormone receptors and the nature of hormone-receptor interactions. *Structure* **3**, 1341–1353 (1995).

15. Kobe, B. & Kajava, A. V. The leucine-rich repeat as a protein recognition motif. *Curr. Opin. Struct. Biol.* **11**, 725–732 (2001).

16. Davis, D., Liu, X. & Segaloff, D. L. Identification of the sites of N-linked glycosylation on the follicle-stimulating hormone (FSH) receptor and assessment of their role in FSH receptor function. *Mol. Endocrinol.* **9**, 159–170 (1995).

17. Bhowmick, N., Huang, J., Puett, D., Isaacs, N. W. & Laphorn, A. J. Determination of residues important in hormone binding to the extracellular domain of the luteinizing hormone/chorionic gonadotropin receptor by site-directed mutagenesis and modeling. *Mol. Endocrinol.* **10**, 1147–1159 (1996).

18. Liu, C. & Dias, J. A. Long loop residues 33–58 in the human glycoprotein hormone common α subunit contain structural components for subunit heterodimerization and human follitropin-receptor binding. *Arch. Biochem. Biophys.* **329**, 127–135 (1996).

19. Arnold, C. J. *et al.* The human follitropin α-subunit C terminus collaborates with a β-subunit cysteine noose and an α-subunit loop to assemble a receptor-binding domain competent for signal transduction. *Biochemistry* **37**, 1762–1768 (1998).

20. Erickson, L. D. *et al.* Synthetic α-subunit peptides stimulate testosterone production *in vitro* by rat Leydig cells. *Endocrinology* **126**, 2555–2560 (1990).

21. Leinung, M. C., Reed, D. K., McCormick, D. J., Ryan, R. J. & Morris, J. C. Further characterization of the receptor-binding region of the thyroid-stimulating hormone α subunit: a comprehensive synthetic peptide study of the α-subunit 26–46 sequence. *Proc. Natl Acad. Sci. USA* **88**, 9707–9711 (1991).

22. Keutmann, H. T., Mason, K. A., Kitzmann, K. & Ryan, R. J. Role of the β 93–100 determinant loop sequence in receptor binding and biological activity of human luteinizing hormone and chorionic gonadotropin. *Mol. Endocrinol.* **3**, 526–531 (1989).

23. Moyle, W. R. *et al.* Co-evolution of ligand-receptor pairs. *Nature* **368**, 251–255 (1994).

Table 1 Diffraction data and refinement

Data collection	
Space group	C2
Bragg spacings (Å)	25–2.9 (3.0–2.9)
Total observations	85,886
Unique reflections	25,282 (2,494)
Completeness (%)	98.6 (97.7)
Average I/σ(I)	12.4 (3.4)
R _{sym} (%)*	6.9 (38.8)
Refinement (25–2.9 Å, F > 0)	
Reflections (work/test)	24,020/1,262
R _{work} (%)†	21.9 (29.7)
R _{free} (%)†	25.9 (35.6)
Average B-factors (Å ²)	56.3
r.m.s.d.	
Bonds (Å)	0.010
Angles (°)	1.427
B-factors (Å ²)	
Main-chain bonds/angles	1.95/2.88
Side-chain bonds/angles	2.90/4.23

Values in parentheses indicate the corresponding statistics in the highest-resolution shell.
**R*_{sym} = (Σ|*I*_{*h*} − ⟨*I*_{*h*}⟩)/Σ*I*_{*h*}, where ⟨*I*_{*h*}⟩ is the average intensity over symmetry equivalent.
†*R*_{work} = Σ||*F*_o − |*F*_c||/Σ|*F*_o|. *R*_{free} is equivalent to *R*_{work}, but calculated for a randomly chosen 5% of reflections, which were omitted from the refinement process.

24. Dias, J. A., Zhang, Y. & Liu, X. Receptor binding and functional properties of chimeric human follitropin prepared by an exchange between a small hydrophilic intercysteine loop of human follitropin and human lutropin. *J. Biol. Chem.* **269**, 25289–25294 (1994).
25. Grossmann, M. *et al.* Substitution of the seat-belt region of the thyroid-stimulating hormone (TSH) β -subunit with the corresponding regions of choriogonadotropin or follitropin confers luteotropic but not follitropic activity to chimeric TSH. *J. Biol. Chem.* **272**, 15532–15540 (1997).
26. Smits, G. *et al.* Glycoprotein hormone receptors: determinants in leucine-rich repeats responsible for ligand specificity. *EMBO J.* **22**, 2692–2703 (2003).
27. Vischer, H. F., Granneman, J. C. M. & Bogerd, J. Opposite contribution of two ligand-selective determinants in the N-terminal hormone-binding exodomain of human gonadotropin receptors. *Mol. Endocrinol.* **17**, 1972–1981 (2003).
28. Chen, F., Wang, Y. & Puett, D. The carboxy-terminal region of the glycoprotein hormone α -subunit: contributions to receptor binding and signaling in human chorionic gonadotropin. *Mol. Endocrinol.* **6**, 914–919 (1992).
29. Strader, C. D., Fong, T. M., Tota, M. R., Underwood, D. & Dixon, R. A. Structures and function of G protein-coupled receptors. *Annu. Rev. Biochem.* **63**, 101–132 (1994).
30. Palczewski, K. *et al.* Crystal structure of rhodopsin: a G protein-coupled receptor. *Nature* **289**, 739–745 (2000).
31. Remy, J. *et al.* Mapping of HCG-receptor complexes. *Mol. Cell. Endocrinol.* **125**, 79–91 (1996).
32. Braun, T., Schofield, P. R. & Sprengel, R. Amino-terminal leucine-rich repeats in gonadotropin receptors determine hormone selectivity. *EMBO J.* **10**, 1885–1890 (1991).
33. Vassart, G., Pardo, L. & Costagliola, S. A molecular dissection of the glycoprotein hormone receptors. *Trends Biochem. Sci.* **29**, 119–120 (2004).
34. Kunishima, N. *et al.* Structural basis of glutamate recognition by a dimeric metabotropic glutamate receptor. *Nature* **407**, 971–977 (2000).
35. Costagliola, S. *et al.* Tyrosine sulfation is required for agonist recognition by glycoprotein hormone receptors. *EMBO J.* **21**, 504–513 (2002).
36. Szkudlinski, M. W., Teh, N. G., Grossmann, M., Tropea, J. E. & Weintraub, B. D. Engineering human glycoprotein hormone superactive analogues. *Nature Biotechnol.* **14**, 1257–1263 (1996).
37. Heikoop, J. C. *et al.* Partially deglycosylated human choriogonadotropin, stabilized by intersubunit disulfide bonds, shows full bioactivity. *Eur. J. Biochem.* **253**, 354–356 (1998).
38. Ji, I., Zeng, H. & Ji, T. H. Receptor activation of and signal generation by the lutropin/choriogonadotropin receptor. *J. Biol. Chem.* **268**, 22971–22974 (1993).
39. Terrillon, S. & Bouvier, M. Roles of G-protein-coupled receptor dimerization. From ontogeny to signaling regulation. *EMBO Rep.* **5**, 30–34 (2004).
40. Grasberger, B., Minton, A. P., DeLisi, C. & Metzger, H. Interactions between proteins localized in membranes. *Proc. Natl Acad. Sci. USA* **83**, 6258–6262 (1986).
41. Osuga, Y. *et al.* Co-expression of defective luteinizing hormone receptor fragments partially reconstitutes ligand-induced signal generation. *J. Biol. Chem.* **272**, 25006–25012 (1997).
42. Ji, I. *et al.* Trans-activation of mutant follicle-stimulating hormone receptors selectively generates only one of two hormone signals. *Mol. Endocrinol.* **18**, 968–978 (2004).
43. Tao, Y. X., Johnson, N. B. & Segaloff, D. L. Constitutive and agonist-dependent self-association of the cell surface human lutropin receptor. *J. Biol. Chem.* **279**, 5904–5914 (2004).
44. Roess, D. A. & Smith, S. M. L. Self-association and raft localization of functional luteinizing hormone receptors. *Biol. Reprod.* **69**, 1765–1770 (2003).
45. Otwinowski, Z. & Minor, W. Processing of X-ray diffraction data collected in oscillation mode. *Methods Enzymol.* **276**, 307–326 (1997).
46. Navaza, J. Implementation of molecular replacement in AMoRe. *Acta Crystallogr. D* **57**, 1367–1372 (2001).
47. Collaborative Computational Project, Number 4. The CCP4 suite: Programs for protein crystallography. *Acta Crystallogr. D* **50**, 760–776 (1994).
48. Jones, T. A., Zou, J. Y., Cowan, S. W. & Kjeldgaard, M. Improved methods for building protein models in electron density maps and the location of errors in the models. *Acta Crystallogr. A* **47**, 110–119 (1991).
49. Brünger, A. T. *et al.* Crystallography & NMR system: A new software suite for macromolecular structure determination. *Acta Crystallogr. D* **54**, 905–921 (1998).
50. Laue, T. M., Shah, B. D., Ridgeway, T. M. & Pelletier, S. L. in *Analytical Ultracentrifugation in Biochemistry and Polymer Science* (eds Harding, S. E., Rowe, A. J. & Horton, J. C.) 90–125 (Royal Society of Chemistry, Cambridge, 1992).

Supplementary Information accompanies the paper on www.nature.com/nature.

Acknowledgements We thank J. Lustbader and L. Lobel for discussion and access to the cDNA clones of human FSHR and FSH, which were originally provided by A. Hsueh and W. Moyle; N. Belgado and D. Hildesheim for help with insect cell culture work; T. Laue, R. Olson and E. Gouaux for advice on analytical ultracentrifugation experiments; M. A. Gawinowicz for sequencing and mass spectrometry; R. Henderson and R. Kovall for critical reading of the manuscript; L. Mosyak, J. Cheung, G. Gregorio, V. Grantcharova, D. Garboczi, C. Lusty and R. Jin for discussion; and synchrotron beamline staff for assistance with data collection. Beamline X4A at the National Synchrotron Light Source (NSLS) is supported by the New York Structural Biology Center, and beamline 31-ID at the Advanced Photon Source (APS) is supported by Structural GenomiX, Inc. NSLS and APS are Department of Energy facilities. Work was supported in part by a NIH grant. Q.R.F. is an Agouron Institute fellow of the Jane Coffin Childs Memorial Foundation.

Competing interests statement The authors declare that they have no competing financial interests.

Correspondence and requests for materials should be addressed to W.A.H. (wayne@convex.hhmi.columbia.edu). Atomic coordinates and structure factors have been deposited into the Protein Data Bank under accession code 1XWD.

Role of the proto-oncogene *Pokemon* in cellular transformation and *ARF* repression

Takahiro Maeda^{1,2}, Robin M. Hobbs^{1,2}, Taha Merghoub^{1,2}, Ilhem Guernah^{1,2}, Arthur Zelent³, Carlos Cordon-Cardo², Julie Teruya-Feldstein² & Pier Paolo Pandolfi^{1,2}

¹Cancer Biology and Genetics Program, ²Department of Pathology, Memorial Sloan-Kettering Cancer Center, Sloan-Kettering Institute, 1275 York Avenue, New York, New York 10021, USA

³Leukemia Research Fund Center at the Institute of Cancer Research, Chester Beatty Laboratories, Fulham Road, London SW3 6JB, UK

Aberrant transcriptional repression through chromatin remodelling and histone deacetylation has been postulated to represent a driving force underlying tumorigenesis because histone deacetylase inhibitors have been found to be effective in cancer treatment. However, the molecular mechanisms by which transcriptional derepression would be linked to tumour suppression are poorly understood. Here we identify the transcriptional repressor *Pokemon* (encoded by the *Zbtb7* gene) as a critical factor in oncogenesis. Mouse embryonic fibroblasts lacking *Zbtb7* are completely refractory to oncogene-mediated cellular transformation. Conversely, *Pokemon* overexpression leads to overt oncogenic transformation both *in vitro* and *in vivo* in transgenic mice. *Pokemon* can specifically repress the transcription of the tumour suppressor gene *ARF* through direct binding. We find that *Pokemon* is aberrantly overexpressed in human cancers and that its expression levels predict biological behaviour and clinical outcome. *Pokemon*'s critical role in cellular transformation makes it an attractive target for therapeutic intervention.

The POZ domain (for poxvirus and zinc finger; also known as the BTB domain) is a highly conserved protein–protein interaction domain found in about 250 known human proteins so far (<http://www.ebi.ac.uk/interpro/IEntry?ac=IPR000210>). Members of the POK (POZ and Krüppel) family of proteins contain an amino-terminal POZ domain and a carboxy-terminal DNA-binding domain made of Krüppel-type zinc fingers and can act as potent transcriptional repressors through the recruitment of histone deacetylases (HDACs) and subsequent chromatin remodelling^{1,2}. POK proteins are important in embryonic development, differentiation and oncogenesis^{3–7}. In particular, PLZF (for promyelocytic leukaemia zinc finger)⁸ and BCL6 (for B-cell lymphoma 6)⁹, two members of this POK family, are involved in chromosomal translocations associated with acute promyelocytic leukaemia and non-Hodgkin's lymphoma, respectively.

Pokemon (for POK erythroid myeloid ontogenic factor; also known as LRF¹⁰, OCZF¹¹ and FBI-1 (ref. 12)) is a POK protein family member that has a critical and pleiotropic function in cellular differentiation and can physically interact with other POK family members such as BCL6 (ref. 10). *Zbtb7* inactivation in mouse results in embryonic lethality and impaired cellular differentiation in multiple tissues including the B-cell compartment (T.M. and P.P.P., unpublished data). It could therefore be proposed that perturbation of *Pokemon*'s transcriptional function impairs cellular differentiation, as is often observed in human cancer. In a non-mutually exclusive hypothesis, *Pokemon* could modulate cellular functions/pathways important for oncogenic transformation, thus taking a more direct role in tumorigenesis. To test this hypothesis we therefore assessed the role of *Pokemon* in cancer pathogenesis through a direct genetic approach.

Role of *Pokemon* in oncogenesis

We initially tested whether loss of *Pokemon* function would affect the transforming potential of primary cells. To this end, we performed classic cell growth and transformation assays by taking advantage of early-passage mouse embryonic fibroblast

cells (MEFs). In this experimental setting, combinations of potent oncogenes such as *E1A*+*H-ras*^{V12}, *Myc*+*H-ras*^{V12}, *T-antigen* (*T-Ag*)+*H-ras*^{V12} normally elicit a marked proliferative response in wild-type (WT) MEFs¹³ (Fig. 1a). In addition, although overexpression of a single oncogene does not transform WT MEFs, combinations of oncogenes can induce cellular transformation as revealed by a colony formation assay in soft agar (Fig. 1b). Strikingly, *Zbtb7*^{−/−} MEFs were completely unresponsive to both the proliferative stimulus triggered by these oncogene combinations and their transforming ability, whereas the expression levels of introduced oncogenes were comparable in WT and *Zbtb7*^{−/−} MEFs (Supplementary Fig. 1a). *Pokemon* is therefore required for the growth-promoting and transforming activity of these combinations of oncogenes.

We next determined whether *Pokemon* would display proto-oncogenic activity when coexpressed in combination with other oncogenes. To this end, we infected WT MEFs with retroviruses containing *Myc*, *H-ras*^{V12}, *T-Ag* and/or *Pokemon* genes and generated a growth curve. In this experimental condition, mock-infected control cells failed to proliferate, because cells were plated sparsely and lost cell–cell contact (not shown). As described previously, MEFs infected with *Myc* alone underwent apoptosis¹⁴, whereas expression of oncogenic *ras* resulted in premature senescence¹⁵ (not shown). As a consequence, MEFs infected singly with *H-ras*^{V12} or *Myc* failed to proliferate (Fig. 1c). In contrast, *Pokemon* coexpression blocked oncogene-induced apoptosis and senescence and, as a result, MEFs co-infected with *Pokemon* and *Myc*, *H-ras*^{V12} or *T-Ag* displayed a marked proliferative advantage (Fig. 1c). Furthermore, *Pokemon*-infected Rat1 MycErTM cells¹⁶ were more resistant to *Myc*-induced apoptosis than mock-infected cells (Supplementary Fig. 1c). Strikingly, MEFs co-infected with *T-Ag*, *Myc*, *H-ras*^{V12} and *Pokemon* readily formed colonies when plated in soft agar (Fig. 1d and Supplementary Fig. 1b). *Pokemon* also opposed *E1A*-induced apoptosis¹⁷, thus stimulating the growth of co-infected MEFs. However, *E1A* and *Pokemon* did not transform MEFs when co-infected (not shown).

Pokemon is a specific repressor of ARF

We set out to investigate the mechanisms by which Pokemon would exert its oncogenic activity. It has been reported that the N-terminal POZ domain of POK family members can mediate their physical interaction with a large co-repressor complex^{1,2}, and Pokemon is no exception to this rule (not shown). We therefore proposed that Pokemon might directly repress the expression of a pivotal tumour suppressor whose silencing would result in cellular transformation in combination with the aforementioned oncogenes. We therefore first characterized the putative Pokemon-binding sequence by CAST (for cyclic amplification and selection of targets; see Supplementary Fig. 2a, b) analysis¹⁸ followed by electrophoretic mobility-shift assay (EMSA). Pokemon directly and specifically bound oligonucleotides containing an identified Pokemon-binding

sequence *in vitro* (Supplementary Fig. 2c). Furthermore, Pokemon could repress the activity of an artificial promoter containing Pokemon-binding sites (Supplementary Fig. 2d). To identify the core binding sequence in a Pokemon-binding site, we generated a series of mutated oligonucleotides and performed EMSA assays. Two cytosine residues in the centre of consensus sequence (underlined in Fig. 2a) were revealed to be essential for Pokemon binding (Supplementary Fig. 2e–g).

Subsequently, we searched the promoters of key tumour suppressor genes for putative Pokemon-binding sites. In this respect, *p19^{Arf}* and *p53* were legitimate candidates in view of the fact that their inactivation is known to cooperate with both *Myc* and oncogenic *ras* in neoplastic transformation¹⁴. We found several putative Pokemon-binding sites in the *ARF* promoter, and several

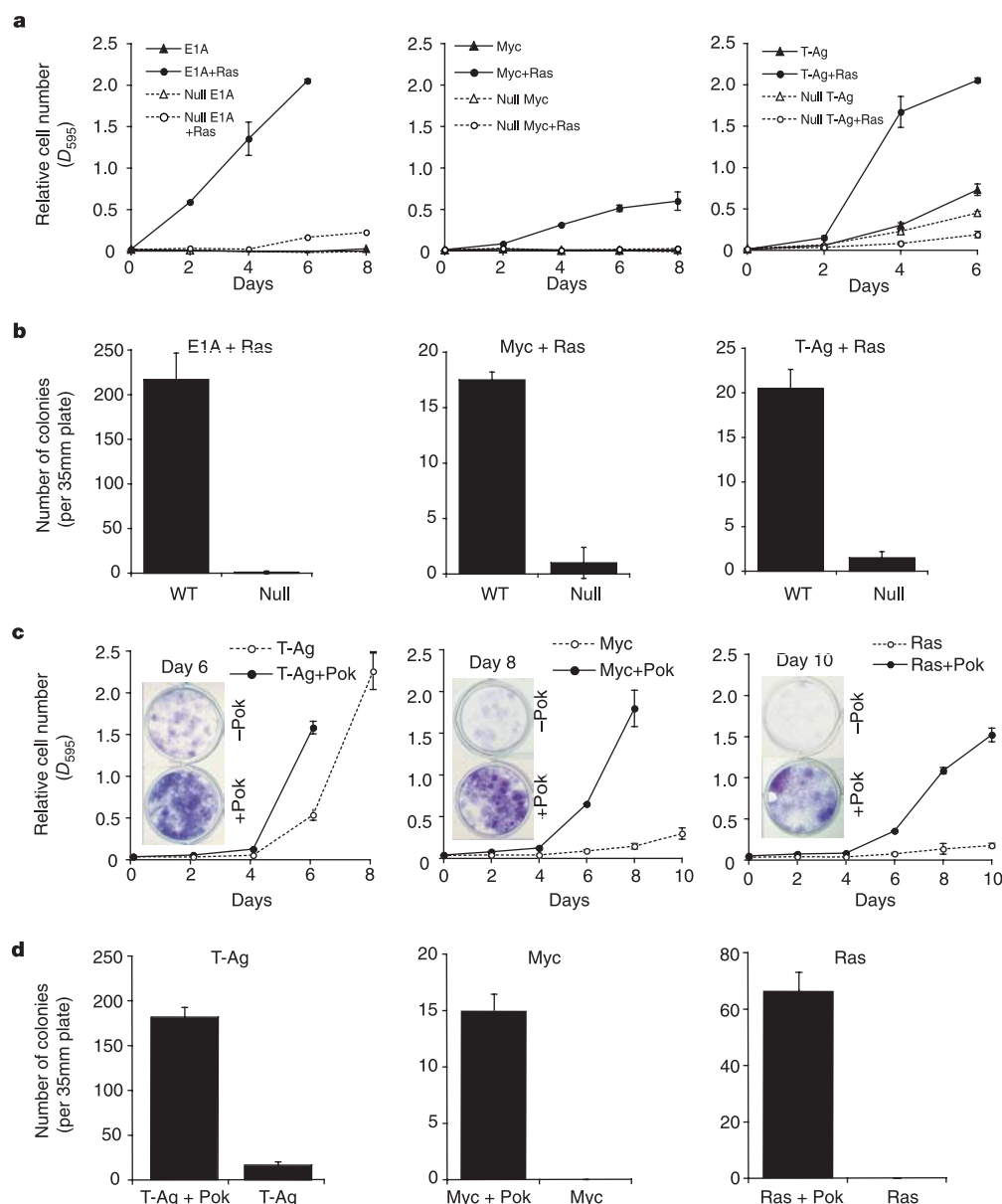


Figure 1 Pokemon is indispensable for cellular transformation and acts as a proto-oncogene. **a**, Growth curve of WT or *Zbtb7*^{-/-} (Null) MEFs infected with the indicated retroviral vectors. After infection, 3,000 cells were plated per well in 12-well plates in triplicate and the cells were fixed on the indicated days for subsequent staining with crystal violet. The y-axis represents relative cell numbers measured by attenuation (*D*) at 595 nm. **b**, Transformation assays in WT or *Zbtb7*^{-/-} (Null) MEFs. Oncogene co-infected MEFs were seeded onto soft agar plates and numbers of transformed foci were scored two

or three weeks later. **c**, Growth curve of WT MEFs infected with retroviral vectors expressing Pokemon (Pok) along with the indicated oncogenes. Growth curves were generated as described in **a**. Representative pictures of wells stained on the indicated days of culture are also shown. **d**, WT MEFs infected with retroviral vectors expressing Pokemon (Pok) along with the indicated oncogenes were plated onto soft agar plates and transformation assays were performed as in **b**. Error bars in **a–d** represent the standard deviation of triplicate wells.

transcriptional modulators, such as E2F1 (refs 19, 20), Bmi-1 (refs 21–26), TBX2, TBX3 (refs 27–29) and Dmp1 (ref. 30), have been reported to regulate the *ARF* promoter (Fig. 2b and Supplementary Fig. 3a).

To determine whether Pokemon would regulate *ARF* gene expression, we first performed transactivation assays with *ARF* promoter constructs fused to luciferase reporters. Pokemon efficiently repressed both *p19^{Arf}* and *p14^{ARF}* basal reporter activity in a

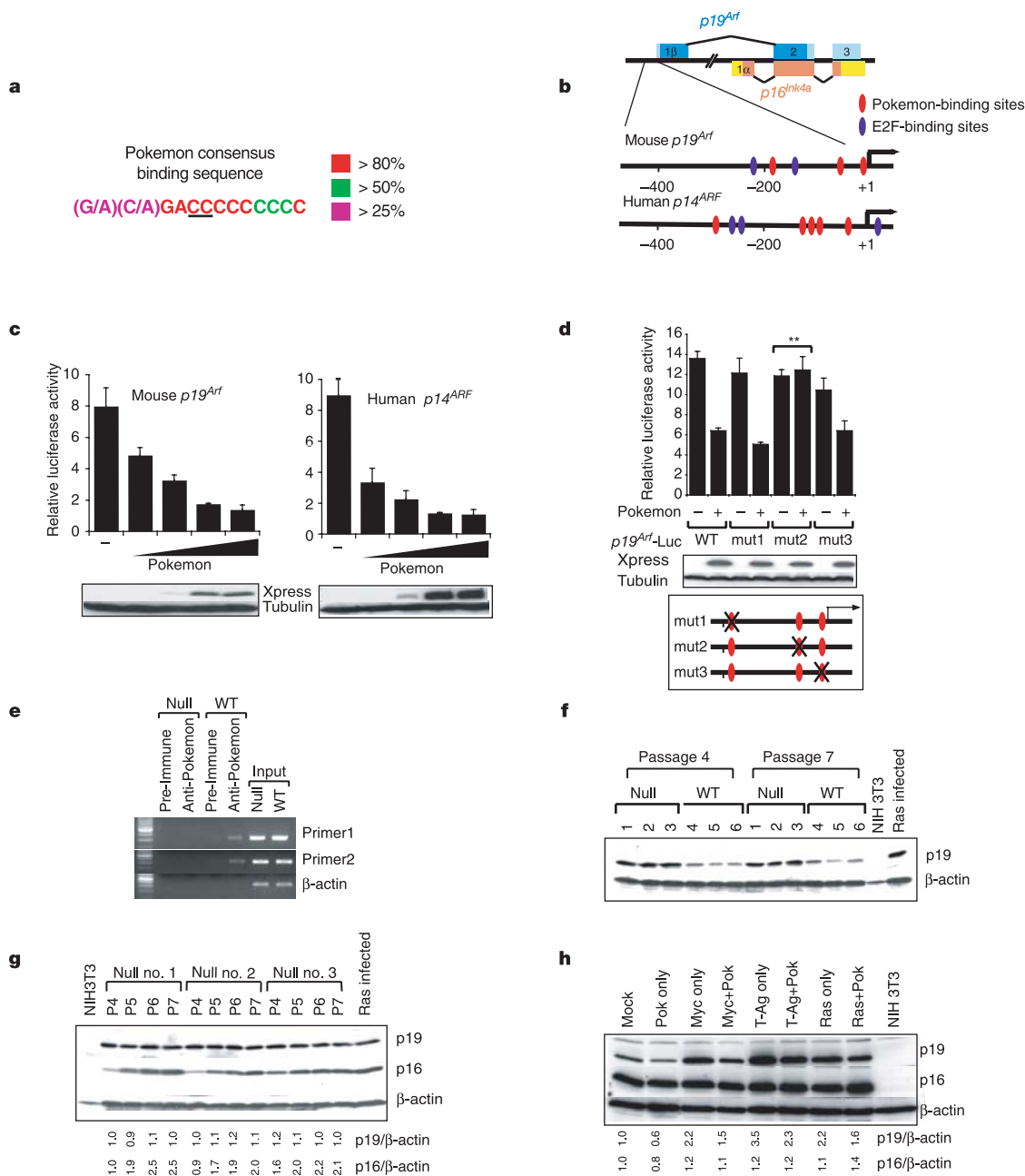


Figure 2 Pokemon is a key *ARF* transcriptional repressor. **a**, Pokemon-binding sequence identified by CAST analysis. The colour code represents the percentage of conserved sequence among pulled-down selected oligonucleotides. The core binding sequence is underlined. **b**, Schematic representation of mouse *p19^{Arf}* and human *p14^{ARF}* genes, promoter regions and the relative positions of putative Pokemon-binding sites. **c**, Pokemon represses both human and mouse *ARF* promoter activity. Transrepression assays in NIH3T3 cells transfected with increasing amounts of Pokemon expression vector along with mouse *p19^{Arf}* and human *p14^{ARF}* luciferase-based reporters. Pokemon expression levels are shown in the bottom panels. **d**, Identification of an essential Pokemon-binding site in mouse *p19^{Arf}* promoter. The putative Pokemon-binding sites in mouse *p19^{Arf}* promoter were mutagenized and subsequently used for reporter assays. Schematic representations of mutated promoter and expression control for the reporter assay are shown (bottom). Pokemon did not repress the promoter activity of mut2 reporter (denoted by asterisk). **e**, ChIP assay in WT and *Zbtb7^{-/-}* (Null) MEFs. Protein–DNA

complexes extracted from MEFs were precipitated either with a preimmune hamster antibody or an anti-Pokemon hamster polyclonal antibody. PCR was performed with primers specific for the *p19^{Arf}* promoter sequence and primers for amplifying mouse β -actin genomic sequence. Amplified DNA was visualized by agarose-gel electrophoresis. **f**, *p19^{Arf}* is aberrantly expressed in *Zbtb7^{-/-}* (Null) MEFs. Western blot analysis of *p19^{Arf}* in serially passaged MEFs. MEFs were prepared from three independent littermates. **g**, Pokemon is a specific *p19^{Arf}* repressor. Western blot analysis of *p19^{Arf}* and *p16^{ink4a}* expression in distinct preparations of serially passaged *Zbtb7^{-/-}* (Null) MEFs. Normalization of *p19^{Arf}* and *p16^{ink4a}* protein levels over β -actin is also shown (bottom). **h**, Pokemon antagonizes oncogene-dependent *p19^{Arf}* induction. Western blot analysis of *p19^{Arf}* and *p16^{ink4a}* protein expression in WT MEFs infected with retroviral vectors expressing the indicated oncogenes. Error bars in **c** and **d** represent the standard deviation of triplicate wells.

dose-dependent manner (Fig. 2c). We also determined that both the POZ and zinc finger domains are required for this repressive activity (Supplementary Fig. 3b). To identify the Pokemon-binding site essential for Pokemon's repressive activity upon the *p19^{Arf}* promoter, we mutagenized the three putative binding sites in reporter constructs and performed luciferase assays (see Supplementary Fig. 3c). As shown in Fig. 2d, Pokemon was unable to repress mut2 reporter activity, which strongly indicates that the Pokemon-binding site located 50 base pairs upstream from the transcription start site is indispensable for *p19^{Arf}* repression. Furthermore, we found that Pokemon efficiently abrogated E2F1-dependent *ARF* transactivation/de-repression in a dose-dependent manner (Supplementary Fig. 3d, and not shown). We next tested whether Pokemon binds the *p19^{Arf}* promoter *in vivo* by chromatin immunoprecipitation (ChIP) assay and designed two different primer sets to amplify the *p19^{Arf}* promoter region (Supplementary Fig. 3a). In these assays, the anti-Pokemon antibody specifically precipitated *p19^{Arf}* promoter sequences from the extract prepared from WT MEFs but not from *Zbtb7^{-/-}* MEFs (Fig. 2e). Furthermore, anti-Pokemon antibody was unable to precipitate β -actin genomic sequences, which indicates that Pokemon binding might be specific for the *p19^{Arf}* promoter. ChIP assays were also performed in transformed MEFs, confirming that Pokemon also binds to the *p19^{Arf}* promoter under transformed conditions (not shown). Thus, Pokemon can bind to the *p19^{Arf}* promoter *in vivo* and repress its activity. The effect of *Zbtb7* inactivation on *p19^{Arf}* expression was then

evaluated. *p19^{Arf}* protein levels were remarkably higher in *Zbtb7^{-/-}* MEFs starting from passage four, when *p19^{Arf}* is normally induced due to culture shock¹⁴ (Fig. 2f). *p19^{Arf}* upregulation also occurred at the mRNA level as indicated by real-time polymerase chain reaction (PCR) analysis (Supplementary Fig. 3e). The *Ink4a-Arf* locus encodes two tumour suppressors, *p16^{Ink4a}* and *p19^{Arf}*, which are homologous to human *p16^{INK4A}* and *p14^{ARF}*, respectively³¹. Both genes are able to induce cellular senescence in response to oncogenic stimuli, and their functional loss is frequently observed in human cancers^{32,33}. Interestingly, in *Zbtb7^{-/-}* MEFs, only *p19^{Arf}* but not *p16^{Ink4a}* was constitutively upregulated, whereas gradual increases in *p16^{Ink4a}* levels with passaging occurred comparably in *Zbtb7^{-/-}* and WT MEFs (Fig. 2g, and not shown). Furthermore, Pokemon was unable to repress the mouse *p16^{Ink4a}* promoter activity as revealed by luciferase assay (Supplementary Fig. 3f), which also supports the notion that Pokemon is a specific repressor for *p19^{Arf}* but not for *p16^{Ink4a}*. In complete agreement with these findings, *p53* (a key downstream target gene of *p19^{Arf}*) was also found overexpressed in *Zbtb7^{-/-}* MEFs (Supplementary Fig. 3g).

We then examined whether Pokemon overexpression would repress *p19^{Arf}* expression levels and whether this would also occur in transforming conditions when *p19^{Arf}* is induced in primary cells. Indeed, Pokemon overexpression invariably resulted in decreased *p19^{Arf}* levels under these conditions (Fig. 2h). *p19^{Arf}* levels were also examined in protein extracts obtained from transformed colonies in these oncogene combinations. *p19^{Arf}* and Pokemon protein

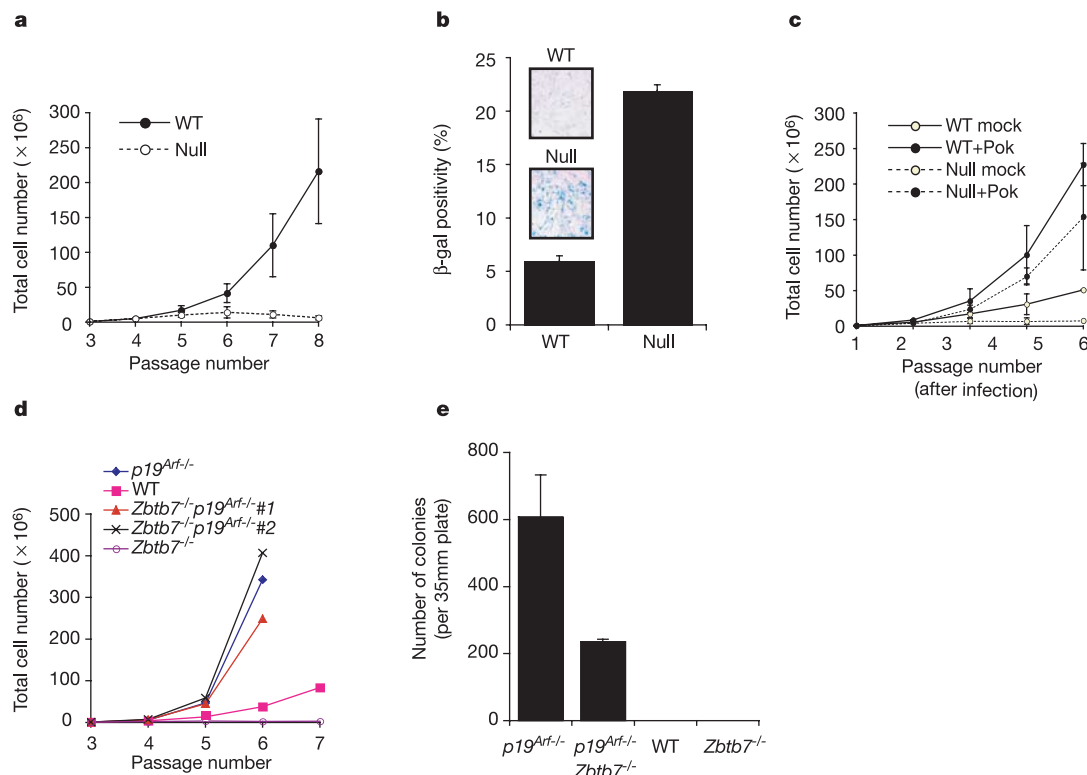


Figure 3 *p19^{Arf}* loss reverts premature senescence and refractoriness to oncogenic transformation in *Zbtb7^{-/-}* MEFs. **a**, Growth curves from WT and *Zbtb7^{-/-}* (Null) MEFs seeded and passaged by following a 3T9 protocol. Each time point represents the mean $\pm$ s.d. of total cumulative cell number from three independent WT and null MEF cultures. **b**, Senescence assays in WT and null MEFs. The y axis represents the percentage of β -galactosidase-positive cells (mean and s.d.) from three independent WT and null MEFs cultures at passage six. A representative result of three independent experiments is shown. **c**, Pokemon add-back rescues the premature senescence phenotype in *Zbtb7^{-/-}* MEFs. Growth curves from WT and *Zbtb7^{-/-}* (Null) MEFs that were either Pokemon-infected or mock-infected. Each time point represents the

mean $\pm$ s.d. of total cumulative cell number from two independent WT and null infected MEF cultures. **d**, *p19^{Arf}* inactivation rescues the premature senescence phenotype in *Zbtb7^{-/-}* MEFs. Growth curves of MEFs generated from embryos obtained from *Zbtb7^{-/-}**p19^{Arf}*^{+/+} intercrosses are shown. MEFs of various genotypes were derived from littermates. Cumulative cell number was counted as described in **a**. **e**, *p19^{Arf}* inactivation overcomes the refractoriness to oncogenic transformation observed in *Zbtb7^{-/-}* MEFs. WT, *Zbtb7^{-/-}*, *p19^{Arf}*^{-/-} and *Zbtb7^{-/-}**p19^{Arf}*^{-/-} MEFs obtained as described in **d** were infected with oncogenic *ras*, and soft agar transformation assays were performed as described previously.

expression levels were inversely correlated, whereas $p16^{\text{Ink4a}}$ expression was not affected (not shown). These data therefore demonstrate that Pokemon is an important and selective repressor of $p19^{\text{Arf}}$ expression.

Loss of *Arf* rescues *Zbtb7*^{-/-} MEF transformability

Accumulation of $p19^{\text{Arf}}$ correlates with the onset of cellular senescence in MEFs after culture shock¹⁴. In full agreement with the finding that Pokemon represses $p19^{\text{Arf}}$ expression and that $p19^{\text{Arf}}$ upregulation is observed in *Zbtb7*^{-/-} MEFs from about passage four, *Zbtb7*^{-/-} MEFs proliferated normally until this passage number and their proliferative capacity rapidly declined at later passages (Fig. 3a). This was accompanied by a marked increase in cellular senescence as demonstrated by staining for the senescence marker β -galactosidase³⁴ (Fig. 3b). As expected, these cell growth defects were fully rescued by Pokemon add-back. In addition, Pokemon overexpression further enhanced the proliferative potential of WT MEFs (Fig. 3c).

We next tested whether $p19^{\text{Arf}}$ loss could rescue the proliferative defect and premature senescence phenotype observed in *Zbtb7*^{-/-} MEFs. To this end, *Zbtb7* and $p19^{\text{Arf}}$ double-knockout MEFs (*Zbtb7*^{-/-} $p19^{\text{Arf}}$ ^{-/-} MEFs) were established from double-mutant embryos. Indeed, the growth defects in *Zbtb7*^{-/-} MEFs were fully reverted by $p19^{\text{Arf}}$ loss, and *Zbtb7*^{-/-} $p19^{\text{Arf}}$ ^{-/-} MEFs proliferated indistinguishably from $p19^{\text{Arf}}$ ^{-/-} MEFs (Fig. 3d). Furthermore, $p19^{\text{Arf}}$ inactivation also reverted the refractoriness to oncogenic transformability observed in *Zbtb7*^{-/-} MEFs. *Zbtb7*^{-/-} $p19^{\text{Arf}}$ ^{-/-} MEFs were susceptible to transformation by oncogenic *ras* alone (Fig. 3e).

Pokemon is oncogenic *in vivo*

To determine whether Pokemon would act as a genuine proto-oncogene *in vivo*, we generated a transgenic mouse model in which Pokemon is overexpressed in immature T and B lymphoid lineage cells taking advantage of an *lckE μ* enhancer/promoter transgenic construct³⁵ (Fig. 4a). Positive transgenic founders were identified and the expression levels of the transgene in F₁ mice from the various founders were assessed by semiquantitative polymerase chain reaction with reverse transcription (RT-PCR) from splenocyte RNA (Supplementary Fig. 4a). We selected two of the highest-expressing lines (numbers 16 and 26) and the F₁ and F₂ generations of each line (litters obtained by intercross with negative mice) were used for further study.

Strikingly, both transgenic lines developed aggressive tumours. The hallmark of the disease in these *lckE μ* -Pokemon mice was massive thymic enlargement, accompanied by lymphadenopathy, splenomegaly, hepatomegaly and tumour infiltration into bone marrow (Fig. 4b). Histopathological examinations of the tumours revealed that the thymus and lymph node architecture was completely effaced and showed a starry-sky pattern with tingible-body macrophages. The lymphomas contained a monotonous population of immature lymphoid cells. The spleen showed expansion of the white pulp and effacement and infiltration of the red pulp by tumour cells. The liver showed prominent periportal, lobular and sinusoidal infiltration by tumour cells (Fig. 4c).

To further characterize the phenotype of these lymphomas, we performed flow cytometry analysis. The lymphomas typically showed an immature T-cell immunophenotype (CD3^+ , CD4^{++} , CD8^{++} CD44^{++} , CD25^- ; Fig. 4d). Furthermore, Wright-Giemsa

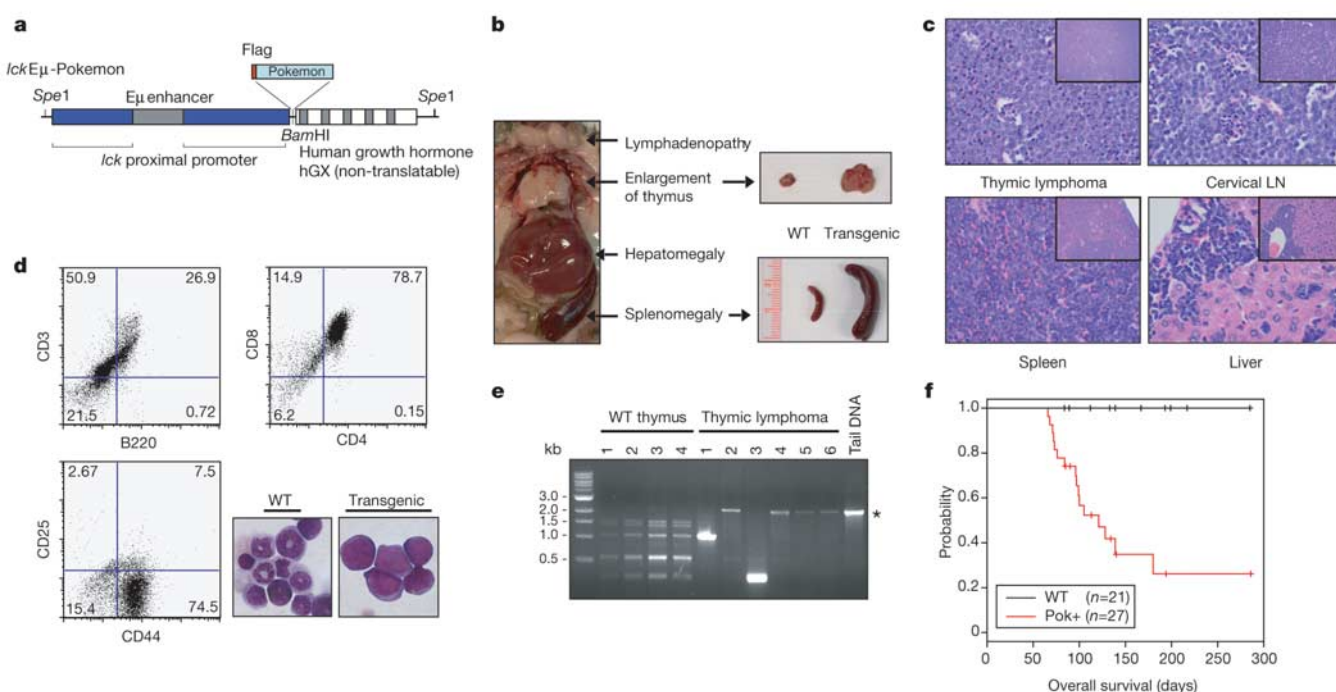


Figure 4 Pokemon transgenic mice develop pre-T LBL. **a**, Diagram of transgenic construct used to overexpress Flag-tagged Pokemon in B lymphocytes and immature T lymphocytes³⁵. **b**, The left panel shows an image of a 3.5-month-old transgene-positive female from the no. 26 line, which showed enlargement of thymus with splenomegaly, hepatomegaly and lymphadenopathy. WT littermate thymus and spleen are compared with enlarged thymus and spleen developed by transgenic mice in the upper and lower right panels respectively. **c**, Sections from an *lckE μ* -Pokemon thymic tumour, enlarged cervical lymph node, spleen and liver stained with haematoxylin and eosin. Infiltration of the spleen and liver can be clearly seen. Both high-power (magnification $\times 400$) and low-power (insets, magnification $\times 40$) views are shown. **d**, Flow cytometry analysis of an *lckE μ* -Pokemon thymic lymphoma for the markers indicated (B220/CD3, CD4/CD8 and

CD44/CD25) showing an immature T-cell immunophenotype (CD3^+ , CD4^{++} , CD8^{++} CD44^{++} , CD25^-). The lower right panels show images of Wright-Giemsa stained cells from bone marrow cytopins of a WT control mouse and a transgene-positive mouse in which the marrow had been completely infiltrated by blasts. **e**, Thymic lymphomas of *lckE μ* -Pokemon mice were monoclonal in origin. The image shows an analysis of D β 1-to-J β 1 rearrangement at the T-cell antigen receptor- β locus by PCR, performed as described³⁹. DNA samples from thymocytes of four WT mice and cells from six *lckE μ* -Pokemon thymic lymphomas were analysed together with a tail DNA control. The asterisk shows the germline configuration. **f**, Kaplan-Meier plot of overall survival of *lckE μ* -Pokemon mice from the no. 26 founder line against WT control littermates.

staining of the bone marrow tumour cells showed large lymphoblasts with round nuclei, fine chromatin and scanty cytoplasm, which is fully consistent with a mouse precursor T-cell lymphoblastic lymphoma/leukaemia (pre-T LBL)³⁶ (Fig. 4d, bottom right). Taking advantage of a specific anti-Pokemon monoclonal antibody 13E9 (Supplementary Fig. 4b), immunohistochemistry of lymphoma sections showed intense nuclear positivity for Pokemon (Supplementary Fig. 4c). However, in normal thymus, Pokemon is expressed mainly in medullary epithelial cells and Hassle's corpuscles, which are negative for the T-cell lineage marker CD3 and positive for the epithelial cell marker cytokeratin AE1:AE3 (Supplementary Fig. 5a, d). In addition, lymphomas were highly positive for terminal deoxynucleotidyl transferase, a diagnostic marker for

lymphoblastic lymphoma/leukaemia (Supplementary Fig. 4c)^{37,38}. The increased Pokemon protein expression in thymic lymphomas was also confirmed by western blot analysis (Supplementary Fig. 4d). To examine whether the lymphomas were monoclonal in origin, we extracted DNA from six different thymic lymphomas and analysed D β 1-to-J β 1 rearrangement of the T-cell antigen receptor- β locus. Four thymic DNA samples from WT mice were used as controls. As shown in Fig. 4e, either a single non-rearranged (or rearranged in other D and J elements) band or a predominant single rearranged band was amplified in lymphoma DNA, whereas PCR products from the four WT thymic DNA demonstrated five possible rearranged and one germline non-rearranged band, as described previously³⁹. These data indicate that the *lckE μ* -Pokemon lympho-

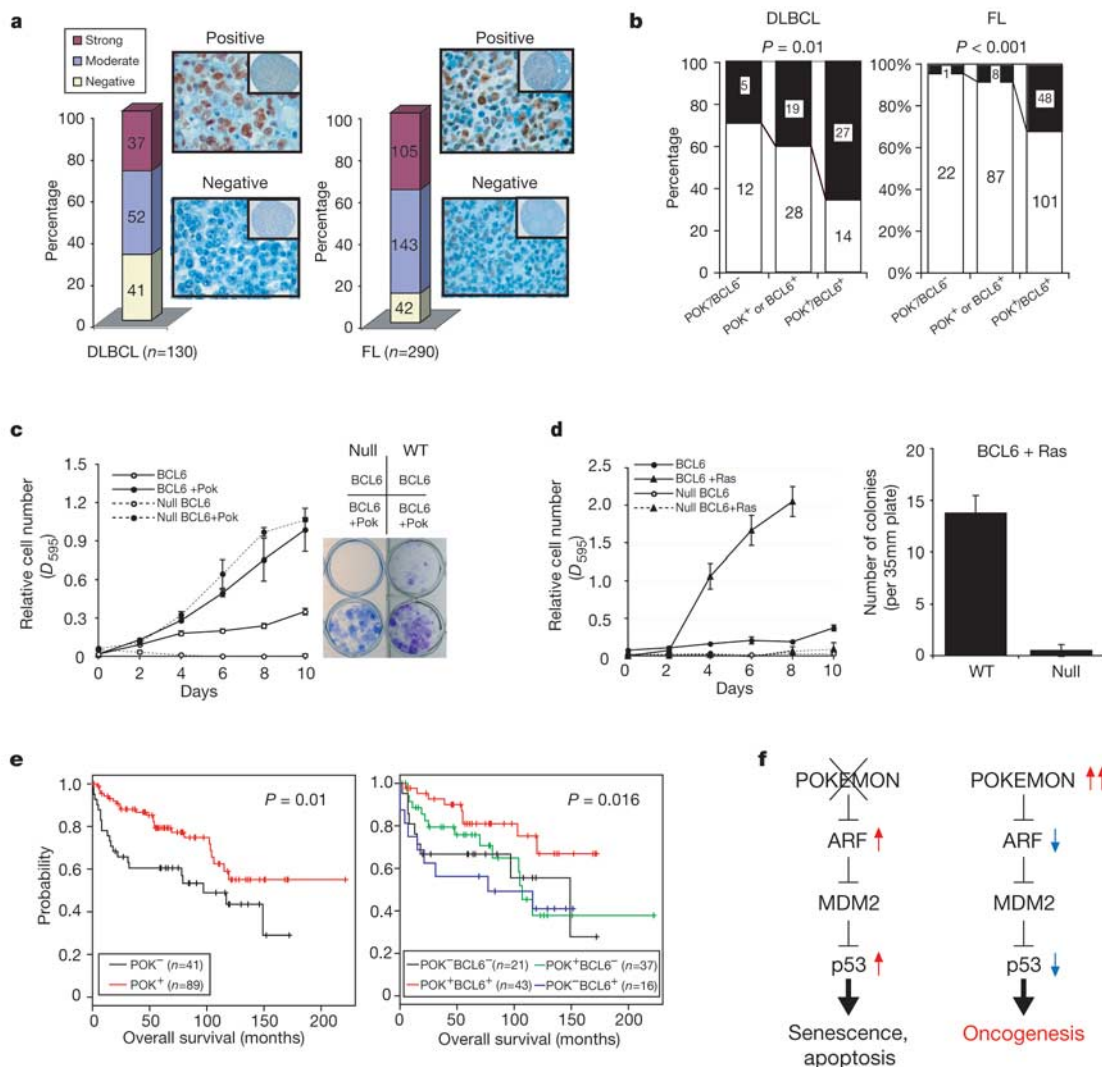


Figure 5 Cooperative roles of Pokemon and BCL6 in lymphomagenesis. **a**, TMA analysis of Pokemon expression in DLBCL and FL clinical samples. Bar graphs show the percentage of Pokemon positivity among samples. Colours indicate levels of Pokemon expression. Absolute numbers of tumours are indicated. Representative pictures of samples positive and negative for Pokemon expression are also shown. **b**, Percentage of Ki-67-positive and Ki-67-negative samples according to Pokemon (Pok) and BCL6 expression pattern. Ki-67-positive and Ki-67-negative samples are represented as black and white bars, respectively. The y-axis represents the percentage of positive and negative cases among each group. Absolute numbers are also indicated. **c**, Pokemon (Pok) is required for BCL6-dependent cell immortalization. Left, growth curves from WT and *Zbtb7*^{-/-} (Null) MEFs after infection with a *BCL6* retroviral vector. Relative cell numbers were measured as described in Fig. 1a. Right, representative pictures of cell cultures stained with crystal violet on day 10. *BCL6* and Pokemon did not confer full transforming

capability to MEFs (not shown). **d**, Left, growth curves from WT and *Zbtb7*^{-/-} (Null) MEFs infected with either *BCL6* alone or *BCL6* together with oncogenic *ras*. Right, transformation assays in soft agar in WT and *Zbtb7*^{-/-} (Null) MEFs. *BCL6* and *H-ras*^{V12} co-infected MEFs of the indicated genotypes were plated onto soft agar plates and the numbers of transformed foci were scored after three weeks. Error bars in **c** and **d** represent the standard deviation of triplicate wells. **e**, Pokemon expression predicts better clinical outcome in DLBCL. Left, overall survival Kaplan-Meier curves of DLBCL cases stratified according to Pokemon-positive and Pokemon-negative expression. Numbers of cases are also indicated. Right, overall survival Kaplan-Meier curves of DLBCL cases stratified according to Pokemon and BCL6 coexpression patterns. Numbers of cases are also indicated. **f**, Proposed model for the role of Pokemon in oncogenesis. Abrogation of Pokemon function leads to cell cycle arrest, cellular senescence and apoptosis, and hence could be exploited for cancer therapy.

mas were monoclonal. These lymphomas were fatal. Specifically, 16 of 27 transgene-positive mice from the number 26 line (the line expressing Pokemon most) developed tumours, and these mice either died or became terminally ill (and hence were killed) between 9 and 40 weeks of age (median 14 weeks) (Fig. 4f). Transgenic mice from the number 16 line also gave rise to pre-T LBL, but at a lower penetrance (4 of 18 mice).

Aberrant Pokemon expression in human cancer

As Pokemon expression proved essential for cellular transformation and this transcription factor is oncogenic when overexpressed both *in vitro* and *in vivo*, we analysed the levels of Pokemon expression in human cancers performing tissue microarray analysis (TMA)⁴⁰ with an anti-Pokemon monoclonal antibody 13E9. We first analysed human T-cell and B-cell lymphomas because *lckEμ*-Pokemon mice developed T-cell lymphomas and Pokemon is normally coexpressed with BCL6 in the B-cell within the germinal centre (Supplementary Fig. 5b, c, and not shown), and might therefore cooperate with BCL6 in lymphomagenesis. We initially performed TMA on small cohorts of T-cell and B-cell lymphoma cases (Supplementary Fig. 6a, b). Interestingly, Pokemon is highly expressed in a subset of T-cell lymphomas, although Pokemon was barely detectable in normal human thymic CD3-positive cells. Furthermore, Pokemon is strongly expressed in diffuse large B-cell lymphoma (DLBCL) and follicular lymphoma (FL), in whose pathogenesis BCL6 is known to be involved.

We subsequently analysed and scored 130 cases of DLBCL and 290 cases of FL for Pokemon protein expression levels (for patient's clinical and immunohistochemical features, see Supplementary Tables 1 and 2 and Supplementary Fig. 6c). Pokemon was highly expressed in 25–35% and moderately expressed in 40–50% of both DLBCL and FL clinical samples (Fig. 5a).

With the use of DNA microarray and TMA analysis, it has recently been proposed that DLBCL can be divided into prognostically significant subgroups with a germinal-centre B-cell-like (GCB) and an activated B-cell-like phenotype^{41–43}. We therefore scored and analysed the expression of GCB markers in our DLBCL cohort and correlated these with Pokemon expression levels. Pokemon expression was not significantly correlated with the GCB phenotype, although Pokemon is expressed in the germinal centre in normal reactive tonsil (Supplementary Table 1).

Interestingly, the vast majority of BCL6-positive DLBCL and FL were found to be Pokemon positive (Supplementary Fig. 6d). In addition, lymphomas positive for both proteins displayed a higher proliferative index than single-negative or double-negative tumours, as revealed by Ki-67 staining, indicating a possible functional cooperation between Pokemon and BCL6 in tumorigenesis (Fig. 5b). BCL6 can immortalize primary cells and fully transform them in cooperation with oncogenic *ras*⁴⁴. Pokemon proved essential for BCL6 immortalizing and transforming activities in MEFs (Fig. 5c, d). Pokemon overexpression not only rescued these defects in *Zbtb7*^{-/-} cells, but also markedly potentiated the growth-promoting activity of BCL6 (Fig. 5c). This cooperative crosstalk between Pokemon and BCL6 in tumorigenesis indicated, in turn, that Pokemon expression and/or Pokemon/BCL6 coexpression might identify a specific lymphoma subtype and possibly predict its clinical outcome. Strikingly, Pokemon positivity was predictive of better overall survival in DLBCL (Fig. 5e, left). Furthermore, Pokemon and BCL6 double positivity predicted an even better clinical outcome for overall survival (Fig. 5e, right).

Discussion

We have demonstrated that Pokemon expression levels are critical determinants of the cellular response to oncogenic transformation. We show that the function of this novel member of the POK family of transcriptional repressors is necessary for oncogenic transformation whereas, conversely, Pokemon can act as a genuine

proto-oncogene when overexpressed (Fig. 5f). Importantly, taking advantage of a transgenic mouse model, we demonstrate that Pokemon can indeed act as a proto-oncogene *in vivo*.

We identify Pokemon as a potent transcriptional repressor of the *p19^{Arf}* tumour suppressor gene. Pokemon is, to our knowledge, the first *ARF*-specific transcriptional repressor to be identified. Furthermore, *p19^{Arf}* inactivation rescued the refractoriness to oncogenic transformation of *Zbtb7*^{-/-} cells. Taken together, these data strongly indicate that, at least in certain cell types, one of the main oncogenic functions of Pokemon consists of its ability to repress *p19^{Arf}* expression. Interestingly, real-time RT-PCR analysis of a cohort of DLBCL cases (*n*=37) revealed that high *Pokemon* gene expression is generally correlated with low expression of the *p14^{ARF}* gene (Supplementary Fig. 6f). Our findings underscore the potential importance of *ARF* suppression in human tumorigenesis.

We show that Pokemon is expressed at very high levels in a subset of human lymphomas. TMA analyses in breast, lung, colon, prostate and bladder carcinomas also reveal high levels of Pokemon expression in a subset of these tumours (T.M. and P.P.P., unpublished data). In this respect, it is worth noting that the genomic region where the *ZBTB7* gene resides (chromosome 19p13.3) is a hotspot for chromosomal translocations (CGAP; <http://cgap.nci.nih.gov/Chromosomes>). We show that Pokemon expression or its coexpression with BCL6 in lymphoma predicts clinical outcome. This is in keeping with the fact that Pokemon is required for BCL6 oncogenic activity, and that high expression of potent proto-oncogenes such as BCL6 and LMO2 also identifies cohorts of patients who respond better to current treatment modalities^{41,43,45}. Cooperative tumorigenesis between these two proto-oncogenes could be exerted through distinct transcriptional programmes or pathways⁴⁶, but also through co-regulation of common target genes in view of their ability to interact physically. In this respect, it is of interest that BCL6 has recently been reported to repress *p53* expression directly⁴⁷.

Our findings provide a further rational for transcription-based therapeutic modalities and identify *Pokemon* as an important target for therapy on the basis of its key role in oncogenesis. □

Methods

Growth curves, transformation and senescence assays

MEFs were prepared from embryos at 13.5 days post coitus. MEFs at passage one or two were used for all experiments. Growth curves after oncogene infections were generated by seeding 3,000 cells (per well) in 12-well plates in triplicate⁴⁸. For transformation assays, 5×10^4 cells were resuspended in 1.5 ml complete Dulbecco's modified Eagle's medium containing 0.3% agarose and seeded into 35-mm plates containing a 3-ml layer of solidified 0.6% agar in complete medium. Foci were scored two or three weeks later. For growth curves of unmanipulated MEFs, 9×10^5 cells were plated in 10-cm dishes and passaged every fourth day. For senescence assays, β -galactosidase-positive cells were scored at passage six in accordance with the manufacturer's specification (Senescence detection kit; Oncogene). The Rat1 MycErTM cells were kindly provided by G. I. Evan. The mock-infected (pWZL-Hygro) cells and the Pokemon-infected (pWZL-Flag-Pokemon) cells were generated by us. After 24 h of serum deprivation, Myc-induced apoptosis was subsequently triggered by the addition of 100 nM 4-hydroxytamoxifen (Sigma) in medium containing 0.1% serum¹⁶, and apoptotic cells were scored with the trypan blue exclusion method on the indicated days.

Molecular analysis

Western blotting was performed in accordance with the standard protocol with the following antibodies: E1A (M58; NeoMarkers), T-Ag (SV40 Ab-1; NeoMarkers), Myc (no. 06-340; Upstate), Ras (Ab-3; Oncogene), Xpress (Invitrogen), *p19^{Arf}* (Ab-1 PC435 lot no. D18714-1; Oncogene), *p16^{Ink4a}* (M-156; Santa Cruz), *p53* (Ab-7; Oncogene) and BCL6 (N-3; Santa Cruz). Anti-Pokemon hamster monoclonal antibody (clone 13E9) was raised against a peptide derived from the Pokemon N-terminal region (amino acids 1–20, MAGGVDGPIGIPFPDHDSSDI), fully conserved between human and mouse proteins. Loading was assessed by β -actin (A-5316; Sigma), α -tubulin (B-5-1-2; Sigma) or Hsp-90 (no. 610419; BD). CAST analysis and EMSA were performed as described¹⁸. For reporter assays, 10^5 NIH 3T3 cells per well were plated into 12-well plates in triplicate 16 h before transfection, and cells were transfected with Lipofectamine 2000. At 24 h after transfection, cells were assayed with a dual luciferase assay kit (Promega).

Plasmids

Mouse *Zbtb7* complementary DNA was amplified by PCR from C57BL/6 mouse bone marrow total cDNA and subcloned into the pcDNA3.1His vector (containing the Xpress

tag; Invitrogen), pSG5 vector (with Flag sequence; Stratagene), pWZL-Hygro and MSCV-PIG vectors (Puro-IRES-GFP, provided by S. W. Lowe). Pokemon deletion mutant constructs used for reporter assays (pcDNA3.1His-POZ, ΔPOZ, Zn and ΔZn) were generated from the pcDNA3.1His-Pokemon vector. Retroviral vectors, expression vectors and luciferase constructs were gifts from A. Koff (pWZL-Hygro-human-c-Myc and pSG5-T-Ag), R. Bernards (pBabe-Hygro-BCL6), S. W. Lowe (pBabe-puro-H-rasV12 and pWZL-Hygro-E1A), W. Kaelin (pRC-CMV-HA-E2F1); S. W. Hiebert (human *p14^{ARF}* luciferase reporter plasmid), M. F. Roussel (mouse *p19^{Arf}* luciferase reporter plasmid), B. A. Mock (mouse *p16^{INK4a}* luciferase reporter plasmid) and R. M. Perlmutter (p1026X vector). Mutated *p19^{Arf}* luciferase reporters were constructed by using a QuikChange Mutagenesis Kit (Stratagene). pWZL-Hygro-T-Ag was generated from the pSG5-T-Ag vector. pWZL-Hygro-H-rasV12 was generated from the pBabe-puro-H-rasV12 vector. 3POKBB-Luc was constructed by insertion of three copies of the Pokemon-binding sequence (5'-GGTTAAAGACCCCTCCCGAATTCGGATC-3') into a luciferase reporter vector TK-LUC from R. Dalla-Favera.

Generation of *Zbtb7^{-/-}* *p19^{Arf}^{-/-}* mutants

Zbtb7^{+/-} mice and *p19^{Arf}^{+/-}* mice (provided by C. J. Sherr) were crossed to generate *Zbtb7^{+/-}* *p19^{Arf}^{+/-}* double-heterozygous mice, which were subsequently intercrossed to generate *Zbtb7^{-/-}* *p19^{Arf}^{-/-}* embryos and MEFs along with WT and single-mutant littermates.

Generation and characterization of *lckEμ*-Pokemon mice

The Flag-Pokemon transgene was produced by ligating the 1.7-kilobase Flag-Pokemon cDNA sequence (from pSG5 Flag-Pokemon vector) into the *Bam*H1 cloning site of the p1026X vector³⁵. For generating transgenic mice, a purified *Spe*I fragment containing *lckEμ*-Pokemon was injected into fertilized (C57BL/6J × CBA/J) F₂ eggs at the MSKCC Transgenic Mouse Core Facility. Transgenic founders were detected by Southern blotting of tail DNA with a Flag-Pokemon cDNA fragment (generated from pSG5 Flag-Pokemon) as a probe. Positive founders were intercrossed with negative founders to establish F₁ generations, and the colony of two highest copy founder lines (nos 16 and 26, revealed by Southern blotting and RT-PCR) were expanded and used for further characterization. To examine the monoclonality of thymic lymphomas, DNA was extracted from WT thymus and thymic lymphomas and the D-to-J rearrangements of the T-cell antigen receptor-β locus were analysed as described previously, using primers 1 and 4 (ref. 39).

TMA analysis

The study cohort comprised 130 DLBCL and 290 FL biopsies consecutively ascertained at the Memorial Sloan-Kettering Cancer Center (MSKCC) between 1984 and 2000. Patient anonymity was ensured and the study received a waiver by the Institutional Review Board. All DLBCL biopsies were obtained at their first evaluation at MSKCC and all DLBCL patients received an anthracyclin-based chemotherapy regimen. The initial histological diagnosis was based on haematoxylin-eosin staining and immunophenotyping results. Biopsies were reviewed and reclassified histologically in accordance with the World Health Organization classification³⁸. TMAs were constructed and stained in the Pathology and Molecular Cytology Core Facilities as described previously⁴⁹. Further detailed information, including staining conditions and antibody dilutions, is described in Supplementary Methods.

Received 17 September; accepted 18 November 2004; doi:10.1038/nature03203.

- Lin, R. J. *et al.* Role of the histone deacetylase complex in acute promyelocytic leukaemia. *Nature* **391**, 811–814 (1998).
- Melnick, A. *et al.* Critical residues within the BTB domain of PLZF and Bcl-6 modulate interaction with corepressors. *Mol. Cell Biol.* **22**, 1804–1818 (2002).
- Barna, M., Hawe, N., Niswander, L. & Pandolfi, P. P. Plzf regulates limb and axial skeletal patterning. *Nature Genet.* **25**, 166–172 (2000).
- Ye, B. H. *et al.* The BCL-6 proto-oncogene controls germinal-centre formation and Th2-type inflammation. *Nature Genet.* **16**, 161–170 (1997).
- Adhikary, S. *et al.* Miz1 is required for early embryonic development during gastrulation. *Mol. Cell Biol.* **23**, 7648–7657 (2003).
- Carter, M. G. *et al.* Mice deficient in the candidate tumor suppressor gene Hic1 exhibit developmental defects of structures affected in the Miller-Dieker syndrome. *Hum. Mol. Genet.* **9**, 413–419 (2000).
- Chen, W. Y. *et al.* Heterozygous disruption of Hic1 predisposes mice to a gender-dependent spectrum of malignant tumors. *Nature Genet.* **33**, 197–202 (2003).
- Chen, Z. *et al.* Fusion between a novel Kruppel-like zinc finger gene and the retinoic acid receptor-α locus due to a variant t(11;17) translocation associated with acute promyelocytic leukaemia. *EMBO J.* **12**, 1161–1167 (1993).
- Ye, B. H. *et al.* Alterations of a zinc finger-encoding gene, BCL-6, in diffuse large-cell lymphoma. *Science* **262**, 747–750 (1993).
- Davies, J. M. *et al.* Novel BTB/POZ domain zinc-finger protein, LRF, is a potential target of the LAZ-3/BCL-6 oncogene. *Oncogene* **18**, 365–375 (1999).
- Kukita, A. *et al.* Osteoclast-derived zinc finger (OCZF) protein with POZ domain, a possible transcriptional repressor, is involved in osteoclastogenesis. *Blood* **94**, 1987–1997 (1999).
- Pessler, F., Pendergrast, P. S. & Hernandez, N. Purification and characterization of FBI-1, a cellular factor that binds to the human immunodeficiency virus type 1 inducer of short transcripts. *Mol. Cell Biol.* **17**, 3786–3798 (1997).
- Weinberg, R. A. The cat and mouse games that genes, viruses, and cells play. *Cell* **88**, 573–575 (1997).
- Zindy, F. *et al.* Myc signaling via the ARF tumor suppressor regulates p53-dependent apoptosis and immortalization. *Genes Dev.* **12**, 2424–2433 (1998).
- Serrano, M., Lin, A. W., McCurrach, M. E., Beach, D. & Lowe, S. W. Oncogenic ras provokes premature cell senescence associated with accumulation of p53 and p16INK4a. *Cell* **88**, 593–602 (1997).
- Evan, G. I. *et al.* Induction of apoptosis in fibroblasts by c-myc protein. *Cell* **69**, 119–128 (1992).
- de Stanchina, E. *et al.* E1A signaling to p53 involves the p19(ARF) tumor suppressor. *Genes Dev.* **12**,

- 2434–2442 (1998).
- Wright, W. E., Binder, M. & Funk, W. Cyclic amplification and selection of targets (CASTing) for the myogenin consensus binding site. *Mol. Cell Biol.* **11**, 4104–4110 (1991).
- Bates, S. *et al.* p14^{ARF} links the tumour suppressors RB and p53. *Nature* **395**, 124–125 (1998).
- Rowland, B. D. *et al.* E2F transcriptional repressor complexes are critical downstream targets of p19(ARF)/p53-induced proliferative arrest. *Cancer Cell* **2**, 55–65 (2002).
- Jacobs, J. J. L., Kieboom, K., Marino, S., DePinho, R. A. & van Lohuizen, M. The oncogene and Polycomb-group gene *bmi-1* regulates cell proliferation and senescence through the *ink4a* locus. *Nature* **397**, 164–168 (1999).
- Jacobs, J. J. *et al.* Bmi-1 collaborates with c-Myc in tumorigenesis by inhibiting c-Myc-induced apoptosis via INK4a/ARF. *Genes Dev.* **13**, 2678–2690 (1999).
- Smith, K. S. *et al.* Bmi-1 regulation of INK4A-ARF is a downstream requirement for transformation of hematopoietic progenitors by E2a-Pbx1. *Mol. Cell Biol.* **23**, 393–400 (2003).
- Kranc, K. R. *et al.* Transcriptional coactivator Cited2 induces Bmi1 and Mel18 and controls fibroblast proliferation via Ink4a/ARF. *Mol. Cell Biol.* **23**, 7658–7666 (2003).
- Kim, J. H. *et al.* The Bmi-1 oncoprotein is overexpressed in human colorectal cancer and correlates with the reduced p16INK4a/p14ARF proteins. *Cancer Lett.* **203**, 217–224 (2004).
- Vonlanthen, S. *et al.* The bmi-1 oncoprotein is differentially expressed in non-small cell lung cancer and correlates with INK4A-ARF locus expression. *Br. J. Cancer* **84**, 1372–1376 (2001).
- Jacobs, J. J. *et al.* Senescence bypass screen identifies TBX2, which represses Cdkn2a (p19(ARF)) and is amplified in a subset of human breast cancers. *Nature Genet.* **26**, 291–299 (2000).
- Lingbeek, M. E., Jacobs, J. J. & van Lohuizen, M. The T-box repressors TBX2 and TBX3 specifically regulate the tumor suppressor gene p14ARF via a variant T-site in the initiator. *J. Biol. Chem.* **277**, 26120–26127 (2002).
- Brummelkamp, T. R. *et al.* TBX-3, the gene mutated in Ulnar-Mammary Syndrome, is a negative regulator of p19ARF and inhibits senescence. *J. Biol. Chem.* **277**, 6567–6572 (2002).
- Inoue, K., Roussel, M. F. & Sherr, C. J. Induction of ARF tumor suppressor gene expression and cell cycle arrest by transcription factor DMP1. *Proc. Natl Acad. Sci. USA* **96**, 3993–3998 (1999).
- Quelle, D. E., Zindy, F., Ashmun, R. A. & Sherr, C. J. Alternative reading frames of the INK4a tumor suppressor gene encode two unrelated proteins capable of inducing cell cycle arrest. *Cell* **83**, 993–1000 (1995).
- Sherr, C. J. & McCormick, F. The RB and p53 pathways in cancer. *Cancer Cell* **2**, 103–112 (2002).
- Lowe, S. W. & Sherr, C. J. Tumor suppression by Ink4a-Arf: progress and puzzles. *Curr. Opin. Genet. Dev.* **13**, 77–83 (2003).
- Dimri, G. P. *et al.* A biomarker that identifies senescent human cells in culture and in aging skin *in vivo*. *Proc. Natl Acad. Sci. USA* **92**, 9363–9367 (1995).
- Iritani, B. M., Forbush, K. A., Farrar, M. A. & Perlmutter, R. M. Control of B cell development by Ras-mediated activation of Raf. *EMBO J.* **16**, 7019–7031 (1997).
- Morse, H. C. III *et al.* Bethesda proposals for classification of lymphoid neoplasms in mice. *Blood* **100**, 246–258 (2002).
- Donlon, J. A., Jaffe, E. S. & Braylan, R. C. Terminal deoxynucleotidyl transferase activity in malignant lymphomas. *N. Engl. J. Med.* **297**, 461–464 (1977).
- Jaffe, E. S., Harris, N. L., Stein, H. & Vardiman, J. W. *Pathology and Genetics of Tumours of Haematopoietic and Lymphoid Tissues* (IARC Press, Lyon, France, 2001).
- Whitehurst, C. E., Chattopadhyay, S. & Chen, J. Control of V(D)J recombinational accessibility of the D beta 1 gene segment at the TCR beta locus by a germline promoter. *Immunity* **10**, 313–322 (1999).
- Gurrieri, C. *et al.* Loss of the tumor suppressor PML in human cancers of multiple histologic origins. *J. Natl Cancer Inst.* **96**, 269–279 (2004).
- Alizadeh, A. A. *et al.* Distinct types of diffuse large B-cell lymphoma identified by gene expression profiling. *Nature* **403**, 503–511 (2000).
- Rosenwald, A. *et al.* The use of molecular profiling to predict survival after chemotherapy for diffuse large-B-cell lymphoma. *N. Engl. J. Med.* **346**, 1937–1947 (2002).
- Hans, C. P. *et al.* Confirmation of the molecular classification of diffuse large B-cell lymphoma by immunohistochemistry using a tissue microarray. *Blood* **103**, 275–282 (2004).
- Shvarts, A. *et al.* A senescence rescue screen identifies BCL6 as an inhibitor of anti-proliferative p19^{ARF}-p53 signaling. *Genes Dev.* **16**, 681–686 (2002).
- Lossos, I. S. *et al.* Prediction of survival in diffuse large-B-cell lymphoma based on the expression of six genes. *N. Engl. J. Med.* **350**, 1828–1837 (2004).
- Sanchez-Beato, M., Sanchez-Aguilera, A. & Piris, M. A. Cell cycle deregulation in B-cell lymphomas. *Blood* **101**, 1220–1235 (2003).
- Phan, R. T. & Dalla-Favera, R. The BCL6 proto-oncogene suppresses p53 expression in germinal-center B cells. *Nature* **432**, 635–639 (2004).
- Carnero, A., Hudson, J. D., Price, C. M. & Beach, D. H. p16INK4A and p19ARF act in overlapping pathways in cellular immortalization. *Nature Cell Biol.* **2**, 148–155 (2000).
- Hedvat, C. V. *et al.* Application of tissue microarray technology to the study of non-Hodgkin's and Hodgkin's lymphoma. *Hum. Pathol.* **33**, 968–974 (2002).

Supplementary Information accompanies the paper on www.nature.com/nature.

Acknowledgements We thank D. Yao, C. Hedvat, M. Dudas, J. Qin and A. Wilton for assistance on TMA preparation, staining and statistical analyses; S. Hasan for assistance with data input and management; K. Manova, C. Farrell and other Molecular Cytology Core Facility members for advice and assistance with IHC; J. Overholser and other Monoclonal Antibody Core Facility staff for help with antibody generation; G. Cattoretti and R. Dalla-Favera for advice; and L. Handker, L. Dong, M. Hu, L. DiSantis and other P.P.P. laboratory members for assistance and discussion. This work is supported in part by an NCI grant to P.P.P.

Competing interests statement The authors declare that they have no competing financial interests.

Correspondence and requests for materials should be addressed to P.P.P. (p-pandolfi@ski.mskcc.org).

A dynamical calibration of the mass–luminosity relation at very low stellar masses and young ages

Laird M. Close¹, Rainer Lenzen², Jose C. Guirado³, Eric L. Nielsen¹, Eric E. Mamajek¹, Wolfgang Brandner², Markus Hartung⁴, Chris Lidman⁴ & Beth Biller¹

¹Steward Observatory, University of Arizona, Tucson, Arizona 85721, USA

²Max-Planck-Institut für Astronomie, Königstuhl 17, D-69117 Heidelberg, Germany

³Departament d'Astronomia i Astrofísica, Universitat de València, E-46100 Burjassot, Valencia, Spain

⁴European Southern Observatory, Alonso de Cordova 3107, Santiago 19, Chile

Mass is the most fundamental parameter of a star, yet it is also one of the most difficult to measure directly. In general, astronomers estimate stellar masses by determining the luminosity and using the 'mass–luminosity' relationship^{1,2}, but this relationship has never been accurately calibrated for young, low-mass stars and brown dwarfs³. Masses for these low-mass objects are therefore constrained only by theoretical models^{1,2}. A new high-contrast adaptive optics camera^{4–6} enabled the discovery of a young (50 million years) companion only 0.156 arcseconds (2.3 AU) from the more luminous (>120 times brighter) star AB Doradus A. Here we report a dynamical

determination of the mass of the newly resolved low-mass companion AB Dor C, whose mass is 0.090 ± 0.005 solar masses. Given its measured 1–2-micrometre luminosity, we have found that the standard mass–luminosity relations^{1,2} overestimate the near-infrared luminosity of such objects by about a factor of ~ 2.5 at young ages. The young, cool objects hitherto thought to be substellar in mass are therefore about twice as massive, which means that the frequency of brown dwarfs and planetary mass objects in young stellar clusters has been overestimated.

Low mass objects are most self-luminous in their youth as they contract and cool. Hence, young objects offer the best opportunity for the direct detection of very-low-mass objects. There have been many efforts to identify young nearby stars to search for planetary mass companions. The K1 star AB Doradus A has long been of great interest owing to its proximity to the Sun (14.9 pc; ref. 7) and its very young age (~ 50 million years, Myr; ref. 8). AB Dor A is considered to be a pre-main-sequence (PMS) star owing to its very fast rotation period (12.3 h; ref. 9), its very strong chromospheric, ultraviolet and X-ray activities¹⁰, and its very high lithium abundance¹¹.

An age of 50 Myr is reasonable for AB Dor A because the strong lithium 6,708-Å equivalent width of AB Dor A¹¹ is similar to K stars in the 50-Myr-old cluster IC 2602 (ref. 12). Moreover, AB Dor is a member of the 'AB Dor moving group', which has an age of 50 Myr (refs 8, 13). AB Dor A is significantly younger than 175 Myr because it is more lithium-rich and rotates more quickly than any similar-temperature member of the 175-Myr-old M35 cluster¹⁴. Moreover, AB Dor A's absolute visual flux is $\sim 200\%$ that of a K1 Pleiades member, as seen in the well-defined ($<10\%$ scatter) M_V versus $m_{Ks} - m_V$ colour magnitude diagram of the 125-Myr-old Pleiades cluster (see figure 4c of ref. 15). However, AB Dor A is probably more than 30 Myr old because the moving group members appear older (less H_α emission) than the 30-Myr-old Tucanae association^{8,13}. Using these independent lines of reasoning, an age of 50^{+50}_{-20} Myr is adopted for AB Dor.

AB Dor is known to have a wide companion AB Dor B^{11,16}. During the commissioning of our high-contrast camera (the new

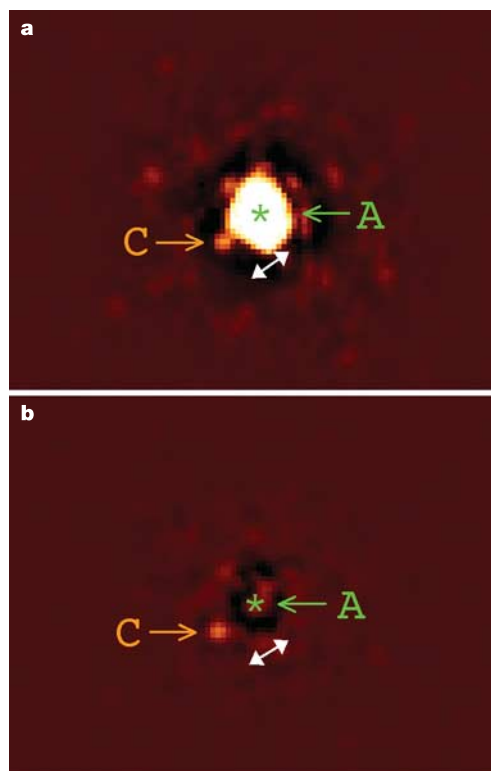


Figure 1 Discovery image of AB Dor C with the NACO SDI high contrast camera. **a**, The 1.625- μm image. **b**, The SDI of the 1.625–1.575- μm images illustrates how the SDI camera enables the subtraction of the scattered light from AB Dor A to reveal the significantly redder source AB Dor C (contrast enhanced in **b**). The white scale bar is the system separation of just 0.156". Both images have had the lowest spatial frequencies removed (unsharp-masked) to help highlight the 'point-like' sources in the field. AB Dor C is the faintest companion ever imaged within 0.16" of a star.

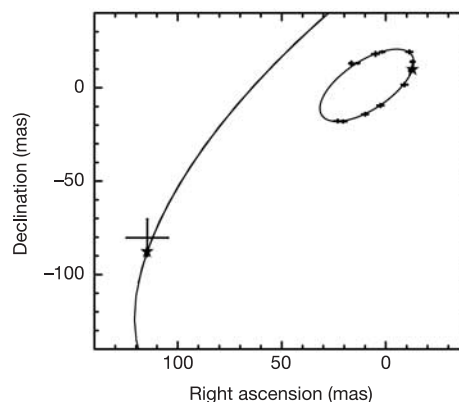


Figure 2 The AB Dor A and C orbital solution. Apparent orbits for AB Dor A's reflex motion (inner ellipse) and AB Dor C (outer ellipse) corresponding to AB Dor A and C masses of $0.865M_\odot$ and $0.090M_\odot$ and a period of 11.75 years. Small data points on the inner ellipse correspond to the VLBI and HIPPARCOS data of ref. 17; the 'star' symbols correspond to the astrometric prediction of the positions of AB Dor A and C on 4 February 2004. The large cross on the outer ellipse corresponds to the latest point from the discovery of AB Dor C. AB Dor A's reflex orbit is fitted by: period $P = 11.75 \pm 0.25$ yr; semi-major axis $a = 0.032 \pm 0.002''$; eccentricity $e = 0.59 \pm 0.03$; periastron passage at 1991.8 ± 0.2 ; inclination $i = 67 \pm 3^\circ$; $\Omega = 132 \pm 2^\circ$; $\omega = 107 \pm 7^\circ$.

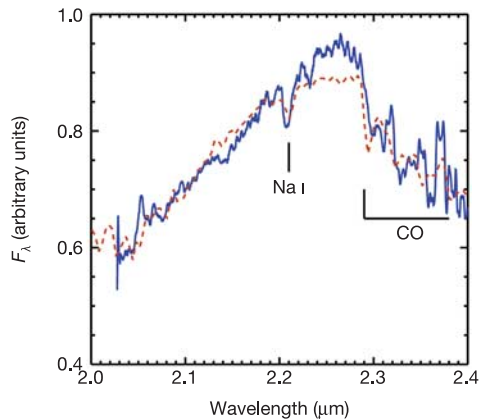


Figure 3 The spectrum of AB Dor C. The K-band spectrum of AB Dor C (solid line) as compared to the best-fit M8 template²⁹ (GL 752B; dashed line). This is the first time a spectrum has been extracted from a companion within $0.16''$ of a primary >80 times brighter. Note the strong $2.2\text{-}\mu\text{m}$ Na I line due to the lower surface gravity of AB Dor C.

NACO Simultaneous Differential Imager, NACO SDI) we obtained observations of AB Dor B with the Very Large Telescope (VLT). We found it to be separated by $9.09 \pm 0.01''$ (135 AU) from AB Dor A at a position angle of $345.9 \pm 0.3^\circ$. It is a common-proper-motion companion of AB Dor A and has a spectral type of M3–M5 (refs 11, 16); it is also known as Rst 137B or HBC 434. We discovered AB Dor B itself to be a tight $0.070''$ binary (henceforth called AB Dor Ba, Bb) with position angle $238.6 \pm 0.3^\circ$ and a Ks flux ratio of $0.67 \pm 0.06\text{ mag}$.

We also detected a faint close companion to AB Dor A (Fig. 1). This companion is the very-low-mass object AB Dor C which had produced the astrometric reflex motion of AB Dor A observed by ref. 17 from very long baseline interferometry (VLBI) and HIPPARCOS data. Encouraged by this result¹⁷, the Hubble Space Telescope with NICMOS tried to detect this faint companion (R. Rebolo, personal communication) but failed owing to the

high contrast required at $<0.3''$ separations. It was not until the commissioning of our SDI camera that direct imaging of AB Dor C occurred.

Once we had identified the exact location of AB Dor C we used the normal imaging modes of NACO to detect the companion at J ($1.2\text{ }\mu\text{m}$), H ($1.65\text{ }\mu\text{m}$), and Ks ($2.1\text{ }\mu\text{m}$). The companion was found to be 80 ± 15 times fainter at Ks, 120 ± 15 times fainter at H and 150 ± 30 times fainter at J (2σ errors). The location of the secondary was $0.156 \pm 0.010\text{ arcsec}$ with a position angle of $127 \pm 1^\circ$ east of north.

To obtain an accurate measurement of the mass of AB Dor C, we used astrometric data¹⁷ to make an exploration of the orbital parameter space of the reflex motion of AB Dor A (similar to the orbit exploration described¹⁷), imposing as a constraint on the search that the predicted relative position of AB Dor A and AB Dor C was coincident, within the uncertainties, with the relative position measured on the NACO SDI detection image (to be conservative, SDI position uncertainties were increased by 50%). Because AB Dor A has $T_{\text{eff}} = 5,081\text{ K}$ and $0.388L_\odot$, where $L_\odot$ is the luminosity of the Sun, calibrated solar-mass PMS models^{1,18} predict masses of $0.853M_\odot$ and $0.849M_\odot$, where $M_\odot$ is the mass of the Sun. In addition, extrapolating from empirical¹⁹ (model-independent) tracks suggests a mass of $0.893M_\odot$. Hence we adopt a mass of $0.865 \pm 0.034M_\odot$ for AB Dor A. Therefore, Kepler's third law applied to the orbital solution (see Fig. 2) yields a mass of $0.090 \pm 0.005M_\odot$ for AB Dor C.

At a mass of just $0.09M_\odot$ AB Dor C is just above the mass of a brown dwarf. Such very-low-mass ($<0.1M_\odot$) objects are very rare close companions ($<1\%$ of stars have brown dwarf companions at $<5\text{ AU}$). At a separation of just 2.3 AU , AB Dor C is the rare example of a very-low-mass object right in the middle of the 'brown dwarf desert'.

From the 2MASS $1\text{--}2.5\text{-}\mu\text{m}$ (near-infrared) fluxes of AB Dor A (which we confirmed from our NACO observations), we derived apparent magnitudes of $m_J = 10.76^{+0.19}_{-0.24}$, $m_H = 10.04^{+0.13}_{-0.15}$ and $m_{Ks} = 9.45^{+0.12}_{-0.15}$ for AB Dor C. Using the HIPPARCOS⁷ parallax distance of 14.94 pc , AB Dor C has absolute magnitudes of $M_J = 9.89^{+0.19}_{-0.24}$, $M_H = 9.19^{+0.13}_{-0.15}$ and $M_{Ks} = 8.57^{+0.12}_{-0.15}$. The near-infrared colours of AB Dor C ($m_J - m_{Ks} = 1.3 \pm 0.4$;

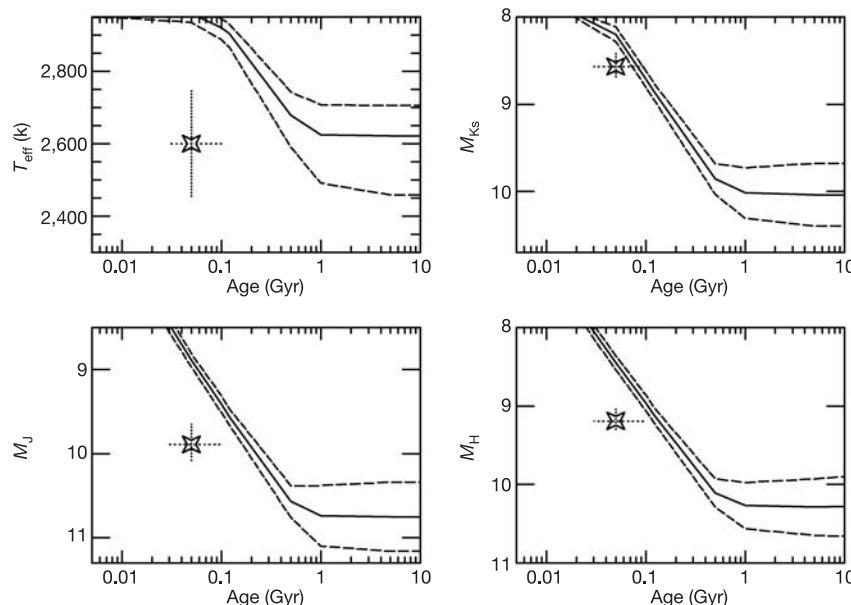


Figure 4 The observations are poorly fitted by the models. The first test of the low-mass DUSTY evolutionary tracks of ref. 1 at young ages (solid line, $0.090M_\odot$ track; dashed lines, $0.085M_\odot$ and $0.095M_\odot$ tracks). Note how throughout the near-infrared (J, H and Ks; $1\text{--}2.5\text{ }\mu\text{m}$) the models systematically overestimate the luminosity and temperature of

AB Dor C (cross shows 2σ errors). Low-mass evolutionary tracks (particularly at H and J) will have to be significantly adjusted to lower luminosities and temperatures to fit the first young mass–luminosity calibrator AB Dor C. The predicted Ks flux is closer than that of H or J.

$m_{\text{H}} - m_{\text{Ks}} = 0.6 \pm 0.3$) are consistent with cool M7–M9 spectral types²⁰.

To estimate better the spectral type of AB Dor C we used the NACO¹⁹ adaptive optics spectrograph. We obtained a total of 20 min of spectra with AB Dor A and C aligned along the 0.086'' NACO slit. We find (see Fig. 3) that AB Dor C is best fitted by a M8 ± 1 spectral template. This suggests an effective temperature²¹ of $2,600 \pm 150$ K. From this temperature and our M_{J} , M_{H} , and M_{Ks} values we estimate²¹ a luminosity of $0.0018 \pm 0.0005 L_{\odot}$ for AB Dor C.

We note at $M_{\text{Ks}} = 8.57$ that AB Dor C is $\sim 200\%$ overluminous compared to known²² late-type (M6.5–M9) Pleiades members at 132 pc, and hence AB Dor C is probably significantly younger than ~ 125 Myr. The 200% overluminosity implies a $\sim 40\%$ larger radius than an M8 object of about the age of the Pleiades (125 Myr). Because AB Dor C would require at least ~ 60 Myr to contract to a Pleiades-sized object², it seems that AB Dor C has an age of ~ 65 Myr. This is consistent with our adopted age of 50^{+50}_{-30} Myr for the system.

From Fig. 4 we see that the standard $0.085 M_{\odot}$, $0.090 M_{\odot}$, and $0.095 M_{\odot}$ DUSTY cooling tracks¹ overestimate the near-infrared luminosity of AB Dor C. Although there is evidence²² that the DUSTY tracks fit older (~ 5 Gyr) low-mass objects it appears that for young objects they overestimate the near-infrared fluxes. This has very significant consequences for the prediction of the mass of young low-mass stars, brown dwarfs, and extrasolar planets. In general, we predict that very young (< 100 Myr), cool (1,700–2,800 K) objects will have masses significantly underestimated by the Next-Gen¹ or DUSTY¹ models. For example, the DUSTY models¹ would suggest that AB Dor C's 50-Myr age and M_{J} , M_{H} and M_{Ks} fluxes are consistent with an $M_{\text{J}} = 0.048 M_{\odot}$, $M_{\text{H}} = 0.052 M_{\odot}$ and $M_{\text{Ks}} = 0.070 M_{\odot}$ brown dwarf, respectively. The predicted mass from the M_{Ks} flux of $0.070 M_{\odot}$ is most accurate, yet is still a significant underestimation. The other popular set of brown-dwarf models² predict a mass of $0.038 M_{\odot}$ for a 50-Myr-old object with $T_{\text{eff}} = 2,600$ K. In general, substellar models^{1,2} predict masses for AB Dor C that underestimate the true $0.090 \pm 0.005 M_{\odot}$ dynamical mass by a factor of two. Such errors will require serious revision of the frequency for the lowest-mass objects (brown dwarfs and extrasolar planets) at young ages.

One direct consequence is that the masses quoted from the DUSTY models for the young objects in deep H and J surveys of the Trapezium Orion²⁴ and σ Orionis²⁵ clusters are not significantly less than the deuterium-burning mass limit ($\sim 0.013 M_{\odot}$) when we lower the luminosity of all the tracks by a constant offset of one magnitude (as required to fit the H and J tracks to the observations of AB Dor C). This has the very significant effect of increasing the mass of the lowest luminosity ($M_{\text{H}} \approx 9.7$ at 1 Myr; ref. 24) members of the Initial Mass Function (IMF) above $0.013 M_{\odot}$. Moreover, the current class of free-floating 'cluster-planets' of $T_{\text{eff}} \approx 1,700$ – $2,800$ K may be misidentified low-mass brown dwarfs (the one 'free-floating planet' too cool ($< 1,200$ K) to apply the DUSTY models for a mass derivation is probably a more-massive-field brown dwarf²⁶). In general, this paper illustrates the danger of using evolutionary tracks at ages that have not been fully calibrated by observations. $\square$

Methods

Simultaneous differential imaging with the new NACO SDI camera

Because of AB Dor's youth and proximity we observed it during commissioning of a new high-contrast adaptive optics camera (NACO SDI^{4–6}). This camera creates four nearly identical simultaneous images of the star and passes this light through three different narrowband methane filters (1.575, 1.600 and 1.625 μm). Because methane has a strong absorption redwards of 1.61 μm , the subtractions of the star's image from 1.575–1.625 μm will remove the star's scattered light halo and reveal any methane-rich planets that were hidden in the glare of the star. This technique of SDI was developed elsewhere²⁷ and has achieved only mixed success before NACO SDI.

To fully capitalize on the power of the SDI technique we^{4,5} developed a double Wollaston beamsplitter which creates four nearly identical images with only 10 nm root-mean-square (r.m.s.) of wavefront error between them. In addition, we developed a very-high-quality glass flat coated with three narrowband 'panel' filters deposited over the four quadrants⁴. When combined with new $\lambda/30$ quality $f/40$ relay optics⁵ (where λ is wavelength and $f/40$ is the focal ratio), we had an optimal SDI camera. This camera is now fully integrated into the facility adaptive optics camera CONICA²⁷ which is fed by the NACO adaptive optics system at the 8-m VLT of the European Southern Observatory (ESO). Our new SDI optics enable a new high-contrast mode for the VLT's NACO²⁸ adaptive optics system. We have found that on bright targets it is possible to reach the theoretical photon-noise limit ($\Delta H = 11$; contrasts of $\sim 25,000$) in 40 min at $0.5''$ at 6σ when using SDI to look for faint methane-rich ($T_{\text{eff}} < 1,200$ K) companions⁶.

During our observations (2 February 2004 UT) of AB Dor we noticed a point-source structure very near the $0.06''$ full-width at half-maximum (FWHM) core of AB Dor (Fig. 1a). Because the SDI camera does not require a coronagraph we can image within $0.1''$ ($2\lambda/D$) of the star, where D is the telescope's diameter. This source was just $0.15''$ from the centre of the star and was not fully subtracted out in the SDI difference images (scattered light, or 'speckles', from the star would normally be strongly attenuated by the SDI pipeline⁶). In particular, the 1.625–1.575- μm difference image from our data reduction pipeline⁶ showed what appeared to be a faint companion (Fig. 1b). The intensity was positive in the 1.625–1.575- μm image, so the point source was not methane-rich (and hence was hotter than 1,200 K) but not as hot as the AB Dor A (spectral type K1V; $T_{\text{eff}} \approx 5,080$ K; ref. 11). We confirmed that the object was a real companion by 'rolling' the VLT's derotator from 0° to 33° and obtaining another SDI data set. As expected for a real companion, the point source indeed rotated with the sky by 33° .

Near-infrared broadband imaging with adaptive optics

Once we had identified the exact location of AB Dor C we used the normal $0.013''$ /pixel imaging modes of NACO to detect the companion at J (1.2 μm), H (1.65 μm), and Ks (2.1 μm). We used an observation pattern similar to that used with the SDI camera. To better enable detectivity we obtained images at derotator rolls of 0° and 180° . The data reduction used the adaptive-optics data reduction pipeline of ref. 20. In addition, we replaced the saturated core of AB Dor A with scaled unsaturated (0.36 s exposure) data from the AB Dor's acquisition image. To enhance the signal from AB Dor C we subtracted the 180° images from the 0° images. Scaled images of the AB Dor A point spread function (PSF) were custom-shifted and subtracted from the 0 – 180° image at the location of C until the companion's signature was removed.

Near-infrared spectra with adaptive optics

To measure the spectral type of AB Dor C we used a $R \approx 1,200$ – $1,500$ (2–2.5 μm) grism with the $0.027''$ per pixel camera of NACO²⁸. We obtained a total of 20 min of spectra with AB Dor A and C aligned along the 0.086'' NACO slit. Eight relatively deep exposures were taken, saturating the core of the AB Dor A spectra, but leaving the wings, containing the signal from AB Dor C, intact. The two objects were nodded along the slit between exposures, and the derotator was rotated by 180° , and the sequence of exposures was repeated. In addition, we obtained four unsaturated spectra of AB Dor A. Strips of unsaturated pixels containing the spectra of AB Dor C were then extracted, wavelength-calibrated, and combined using standard Image Reduction and Analysis Facility (IRAF) routines. Removal of the telluric lines was done via spectra of a G3 standard star (HIP 24153) taken within 30 min of our science exposures, and features from the calibration star were removed using a modified solar spectrum³⁰. Finally, the high signal-to-noise unsaturated AB Dor A observations are subtracted from the extracted spectrum to reveal the spectrum of the companion (see Fig. 3).

Received 26 July; accepted 19 November 2004; doi:10.1038/nature03225.

- Chabrier, G., Baraffe, I., Allard, F. & Hauschildt, P. Evolutionary models for very low-mass stars and brown dwarfs with dusty atmospheres. *Astrophys. J.* **542**, 464–472 (2000).
- Burrows, A., Hubbard, W. B., Lunine, J. I. & Liebert, J. The theory of brown dwarfs and extrasolar giant planets. *Rev. Mod. Phys.* **73**, 719–765 (2001).
- Bouy, H. *et al.* First determination of the dynamical mass of a binary L dwarf. *Astron. Astrophys.* **423**, 341–352 (2004).
- Close, L. M. *et al.* Extrasolar planetary science with adaptive optics. ESO workshop on adaptive optics. *Astron. Soc. Pacif.* (in the press).
- Lenzen, R., Close, L. M., Brandner, W., Biller, B. & Hartung, M. A novel simultaneous differential imager for the direct imaging of giant planets. *SPIE Symp.* **5492**, (in the press).
- Biller, B. *et al.* An algorithm for the suppression of speckle-noise with the VLT & MMT SDI cameras. *SPIE Symp.* **5490**, (in the press).
- Perryman, M. A. C. *et al.* The HIPPARCOS catalogue. *Astron. Astrophys.* **323**, L49–L52 (1997).
- Zuckerman, B., Song, I. & Bessell, M. S. The AB Doradus moving group. *Astrophys. J.* **613**, L65–L68 (2004).
- Innis, J. L., Coates, D. W., Thompson, K. & Robinson, R. D. A study of the rapidly rotating variable star HD 36705 (AB Doradus). *Astron. Soc. Aust.* **6**, 156–160 (1985).
- Mewe, R., Kaastra, J. S., White, S. M. & Pallavicini, R. Simultaneous EUVE & ASCA observations of AB Doradus: temperature structure and abundances of the quiescent corona. *Astron. Astrophys.* **315**, 170–178 (1996).
- Vilhu, O., Gustafsson, B. & Walter, F. M. Spectroscopy of southern active stars. II. HD 32918, HD 82558, BD - 22 deg 3467, AB Doradus (HD 36705) and RST 137 B. *Astron. Astrophys.* **241**, 167–175 (1991).
- Randich, S., Pallavicini, R., Meola, G., Stauffer, J. R. & Balachandran, S. C. Membership, lithium, and metallicity in the young open clusters IC 2602 and IC 2391: Enlarging the sample. *Astron. Astrophys.* **372**, 862–878 (2001).
- Stelzer, B. & Neuhauser, R. X-ray emission from young stars in the Tucanae association. *Astron. Astrophys.* **361**, 581–593 (2000).

14. Barrado y Navascués, D., Deliyannis, C. P. & Stauffer, J. R. WIYN open cluster study. V. Lithium depletion and metallicity in G and K dwarfs of the open cluster M35. *Astrophys. J.* **549**, 452–466 (2001).
15. Stauffer, J. R. *et al.* Why are the K dwarfs in the Pleiades so blue? *Astron. J.* **126**, 833–847 (2003).
16. Martin, E. L. & Brandner, W. On the evolutionary status of two very active visual binaries. *Astron. Astrophys.* **294**, 744–746 (1995).
17. Guirado, J. C. *et al.* Astrometric detection of a low-mass companion orbiting the star AB Doradus. *Astrophys. J.* **490**, 835–846 (1997).
18. D'Antona, F. & Mazzitelli, I. Stellar models and luminosity functions for the Population II Main Sequence down to its lower end. *Astrophys. J.* **456**, 329–349 (1996).
19. Hillenbrand, L. A. & White, R. J. An assessment of dynamical mass constraints on Pre-Main-Sequence evolutionary tracks. *Astrophys. J.* **604**, 741–757 (2004).
20. Close, L. M., Siegler, N., Freed, M. & Biller, B. Detection of nine M8.0–L0.5 binaries: the very low mass binary population and its implications for brown dwarf and very low mass star formation. *Astrophys. J.* **587**, 407–422 (2003).
21. Luhman, K. L. Young low-mass stars and brown dwarfs in IC 348. *Astrophys. J.* **525**, 466–481 (1999).
22. Martín, E. L. *et al.* Membership and multiplicity among very low mass stars and brown dwarfs in the Pleiades cluster. *Astrophys. J.* **543**, 299–312 (2000).
23. Ségransan, D. *et al.* Mass-luminosity relations of very low mass stars. *Proc. IAU Symp.* **211**, 413–420 (2003).
24. Lucas, P. W. & Roche, P. F. A population of very young brown dwarfs and free-floating planets in Orion. *Mon. Not. R. Astron. Soc.* **314**, 858–864 (2000).
25. Zapatero Osorio, M. R. *et al.* The substellar population in σ Orionis. *Proc. IAU Symp.* **211**, 111–115 (2003).
26. Burgasser, A. J. *et al.* S Orionis 70: just a foreground field brown dwarf? *Astrophys. J.* **604**, 827–831 (2004).
27. Marois, C. *et al.* TRIDENT: an infrared camera optimized for the detection of methanated substellar companions of nearby stars. High-contrast imaging for exo-planet detection. (ed. Schultz, A. B.) *Proc. SPIE* **4860**, 130–137 (2003).
28. Lenzen, R. *et al.* NAOS-CONICA first on sky results in a variety of observing modes. Instrument design and performance for optical/infrared ground-based telescopes. (eds Iye, M. & Moorwood, A. F. M.) *Proc. SPIE* **4841**, 944–952 (2003).
29. Wilking, B. A., Green, T. P. & Meyer, M. R. Spectroscopy of brown dwarf candidates in the ρ Ophi molecular core. *Astron. J.* **117**, 469–482 (1999).
30. Maiolino, R., Rieke, G. H. & Rieke, M. J. Correction of the atmospheric transmission in infrared spectroscopy. *Astron. J.* **111**, 537–545 (1996).

Acknowledgements We thank A. Kaufer, M. Kasper, J. Spyromilio, P. Gandhi and R. Gilmozzi for assistance with SDI commissioning. We thank O. Marco for support during the spectral observations. We thank D. McCarthy and M. Rademacher for help with the initial design of the NACO SDI optics. L.M.C. and B.B. acknowledge support from a NASA Origins grant. J.C.G. acknowledges support from a Spanish DGICYT grant. E.E.M. is supported by a Clay Fellowship from the Smithsonian Astrophysical Observatory. L.M.C. acknowledges support from an NSF Career award and SEC and EMC.

Competing interests statement The authors declare that they have no competing financial interests.

Correspondence and requests for materials should be addressed to L.M.C. (lclose@as.arizona.edu).

High-velocity streams of dust originating from Saturn

Sascha Kempf¹, Ralf Srama¹, Mihály Horányi², Marcia Burton³, Stefan Helfert¹, Georg Moragas-Klostermeyer¹, Mou Roy³ & Eberhard Grün^{1,4}

¹MPI für Kernphysik, Saupfercheckweg 1, 69117 Heidelberg, Germany

²Laboratory for Atmospheric and Space Physics, and Department of Physics, University of Colorado, Boulder, Colorado 80309, USA

³Jet Propulsion Laboratory, 4800 Oak Grove Drive, Pasadena, California 91109, USA

⁴Hawaii Institute of Geophysics and Planetology, University of Hawaii, 1680 East West Road, Honolulu 96822, USA

High-velocity submicrometre-sized dust particles expelled from the jovian system have been identified by dust detectors on board several spacecraft^{1,2}. On the basis of periodicities in the dust impact rate, Jupiter's moon Io was found to be the dominant source of the streams³. The grains become positively charged

within the plasma environment of Jupiter's magnetosphere, and gain energy from its co-rotational electric field⁴. Outside the magnetosphere, the dynamics of the grains are governed by the interaction with the interplanetary magnetic field that eventually forms the streams⁵. A similar process was suggested for Saturn⁶. Here we report the discovery by the Cassini spacecraft of bursts of high-velocity dust particles ($\geq 100 \text{ km s}^{-1}$) within ~ 70 million kilometres of Saturn. Most of the particles detected at large distances appear to originate from the outskirts of Saturn's outermost main ring. All bursts of dust impacts detected within 150 Saturn radii are characterized by impact directions markedly different from those measured between the bursts, and they clearly coincide with the spacecraft's traversals through streams of compressed solar wind.

In agreement with model predictions for the interplanetary dust environment between Jupiter and Saturn, the cosmic dust analyser⁷ (CDA) on Cassini registered almost no dust impacts during the late phase of the spacecraft's cruise to Saturn. But beginning in 2004 when Cassini was closer to Saturn than $1,200 R_S$ (Saturn's radius $R_S = 60,330 \text{ km}$), two faint impact bursts by high-velocity grains were observed. Remarkably, the burst intensity seemed to grow as Cassini approached Saturn (see Fig. 1), which indicates a saturnian origin of the impactors. Interplanetary dust can be excluded as the source by examining the transmitted impact signals. At 8.5 AU from the Sun, interplanetary particles moving in bound orbits about the Sun hit the Cassini dust detector with velocities below $10 \text{ km s}^{-1} (1 + \max(e))^{1/2} + \max(v_{sc}) \approx 20 \text{ km s}^{-1}$, where $e < 1$ is the eccentricity and $v_{sc} < 6 \text{ km s}^{-1}$ is the inertial spacecraft velocity. (Here one astronomical unit, AU, is the average Earth–Sun distance.) According to the CDA calibration, impact speeds of about 20 km s^{-1} will lead to rise times of the ion grid signal longer than $18 \mu\text{s}$, which is much longer than the observed rise times of below $3 \mu\text{s}$. In fact, the transmitted signals are steeper than the fastest calibration impact of about 70 km s^{-1} recorded at the Heidelberg dust accelerator facility (Fig. 2). Consequently, the impact speed of the grains comprising the bursts even exceeds 70 km s^{-1} , which excludes bound interplanetary grains as the source. Further, before Cassini's Saturn orbit insertion the CDA detector was geometrically insensitive to fast interplanetary dust in unbound orbits (the so-called β -meteoroids⁸) as well as to interstellar particles. These observations justify our considering the saturnian system as the origin of the observed particles.

There is, however, a striking similarity in both the rise time and the amplitude of the impact signals produced by saturnian dust grains and those produced by jovian stream particles recorded by CDA during Cassini's fly-by of Jupiter in 2000 (Fig. 2), suggesting that the mass and the impact speeds are at least comparable. This implies, however, that the mass and the impact speed of the detected particles are well outside the calibrated range, and that the standard techniques to derive the dust mass and the collision velocity from the impact signals⁹ cannot be applied to them. We benefit from the fact that the characteristic properties of jovian stream particles are known from theoretical considerations. It was shown¹⁰ that such particles had velocities exceeding 200 km s^{-1} and charge-to-mass ratios of $1,000 \text{ C kg}^{-1}$ (corresponding to grain sizes below 10 nm). We therefore conclude that the impact signals registered by Cassini were due to grains with impact speeds in excess of 100 km s^{-1} and with a typical mass below 10^{-21} kg .

The potential mechanism accelerating particles within Jupiter's and Saturn's magnetosphere to such a high velocity has been discussed in great detail in the literature^{5,6,11}, so we simply summarize the basic concept. Positively charged grains will be accelerated outward by the outward-pointing co-rotational electric field caused by Saturn's rotating magnetic field. Possible sources of positively charged dust in Saturn's magnetosphere are the dense A ring outside the synchronous radius at $1.86 R_S$, the E ring, and the

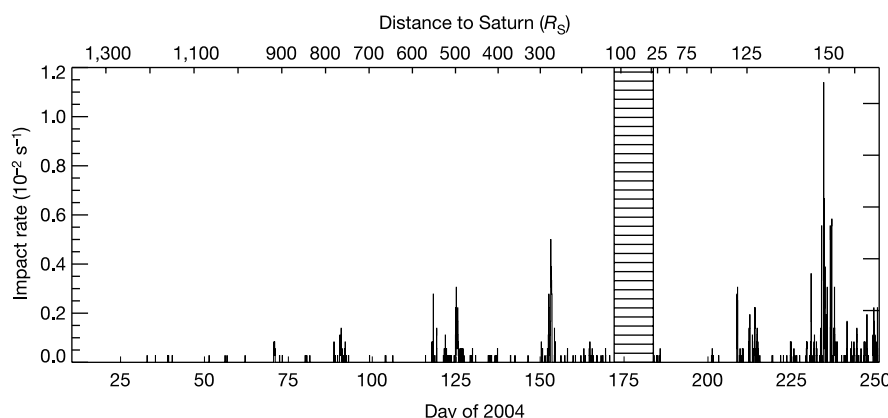


Figure 1 Impact rate registered by the cosmic dust analyser (CDA) between 10 January and 6 September 2004. Between 20 June and 1 July, embracing Cassini's insertion into its orbit about Saturn (Saturn orbit insertion, SOI) the instrument was powered off (marked by the vertical stack of bars). The upper scale gives Cassini's distance to Saturn in R_S . In total, 1,409 impacts were detected (before SOI, 546 impacts; after SOI, 863

impacts). All of them showed the characteristic features of a high-velocity impact by a tiny dust particle. Note that owing to Cassini's highly irregular attitude profile, the observed flux onto the CDA does not necessarily represent entirely the dust environment during this period.

dust clouds around the icy moons Dione and Rhea. To overcome Saturn's gravity, the charge-to-mass ratio Q/m of the grain has to be sufficiently large. This constraint sets an upper limit on the radius of a grain escaping from the inner saturnian system. There is also a lower size limit, as Q/m has to be sufficiently small that grains do not circle along a magnetic field line and become tied to Saturn's magnetic field. Grains of appropriate size leave Saturn's magnetosphere at the speed (in km s^{-1})

$$v_e \approx 35 \left[L^* - \frac{1}{2} \right]^{1/2} \left[\frac{R_S}{R} \right]^{1/2}$$

where $L^* \approx 542 (R_d)^{-2} (\Phi) (1 - [R/20R_S])$ is the ratio of the force due to the co-rotational electric field to the force due to Saturn's gravity, Φ is the average surface potential of the grain (in V), and R_d is the grain radius (in nm). The dust is assumed to consist of water ice (density 10^3 kg m^{-3}). The charge of a grain in space, in general, results from the competition between various charging processes, including the collection of electrons and ions, and the emission of

photo and secondary electrons. Using a model of the plasma parameters in Saturn's magnetosphere based on the Voyager measurements¹², the equilibrium surface potential of grains can be calculated. There are two regions where the surface potential of a dust particle is expected to be positive (lower panel of Fig. 3) and which could be the source regions of the ejected stream particles: (1) at the outskirts of the A ring, where owing to absorption by the rings the plasma density remains very low and photo-emission due to solar ultraviolet radiation dominates charging; and (2) beyond the orbit of the moon Dione at $R = 6.3R_S$. Grains starting from within the second region must be very tiny ($R_d < 3 \text{ nm}$) to gain impact speeds faster than 70 km s^{-1} (upper panel of Fig. 3).

Outside Saturn's magnetosphere, the dynamics of the grains are governed by the interaction with the interplanetary magnetic field (IMF) convected by the solar wind. At large solar distances the azimuthal component of the IMF dominates, and thus the out-of-ecliptic component of the dust velocity should be affected most. As the inertial spacecraft velocity of about 10 km s^{-1} is small compared with the dust speed, the deviation of the impact direction from

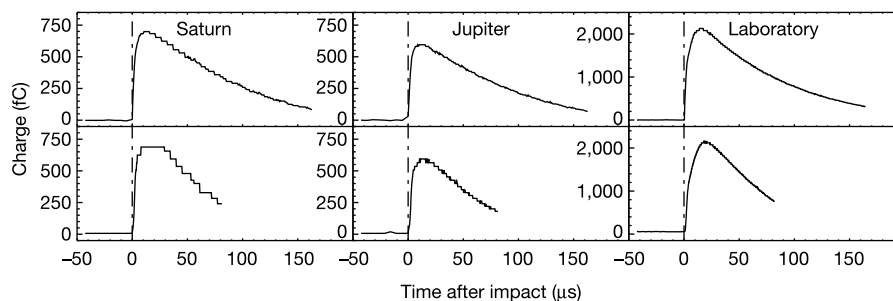


Figure 2 Comparison between typical signals caused by saturnian and jovian stream particles and by the fastest calibration impact. The CDA impact detection is primarily based on the analysis of the plasma generated by the particle impact onto a hemispherical target. After the separation of the plasma constituents within an electric field between the impact target and the ion grid, the evolution of the plasma charges are monitored by low-noise integrating charge amplifiers. In contrast to the CDA's progenitors on the Galileo and Ulysses spacecraft, the full impact signals (rather than a few characteristic parameters) are transmitted to Earth, allowing a highly reliable easy distinction between noise events and genuine impacts. The rise time of an individual

plasma charge signal is a measure of the impact speed, while the total plasma charge depends upon both the impact speed and the dust mass. The upper row shows the evolution of the impact plasma electrons collected at the impact target, while the lower row shows the evolution of the plasma ions collected at the ion grid. The speed of the fastest calibration impact (right column) recorded at the Heidelberg dust accelerator facility was 63 km s^{-1} (mass, $2 \times 10^{-18} \text{ kg}$). Note that the amplitude of the calibration impact is about four times larger than that for the stream particle impacts, which indicates that the dust grains are tiny.

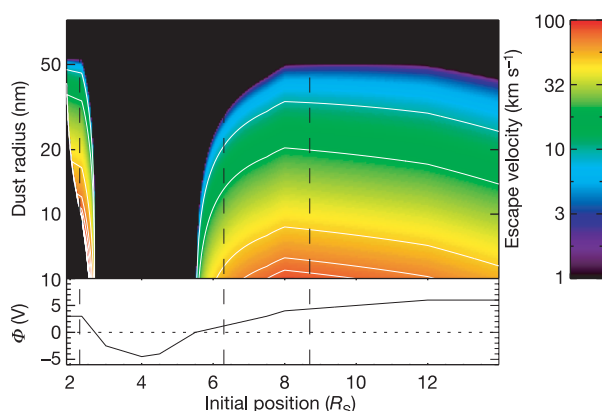


Figure 3 Dependence of the grain's dynamics on its initial position. Top, dependence of the escape velocity upon the initial position and the grain size. The escape velocity is shown as a colour scale. Bottom, the electrostatic equilibrium potential of dust moving within the ring plane⁶ (for calculating the potential, the secondary electron production parameter $\delta_m = 3$ was used, a realistic value for tiny grains¹³). The vertical dashed lines mark, from left to right, the outer edge of the A ring and the location of Saturn's moons Dione and Rhea.

Saturn's location has to be attributed to a bending of the dust trajectories by the Lorentz force due to the IMF. Unfortunately, the spacecraft attitude profile before Saturn orbit insertion was quite complex, occasionally leading to incomplete coverage of the relevant impact directions. Detailed analysis of the dust directionality was only possible between 18 July and 6 September 2004 (Fig. 4). During this period we identified three pronounced impact bursts. Impacts registered between the bursts generally arrived from directions close to the line-of-sight to Saturn, indicating that the dynamics of these particles was only weakly affected by the solar wind. During the bursts, however, the directionality of the dust changed almost instantaneously by about 90° (grains arrive by about 45° out of Saturn's ring plane). The bursts occurred when the Cassini magnetometer observed 'co-rotating interaction regions' (CIRs) (M. Dougherty, personal communication)—compressed, high-speed solar wind characterized by enhanced magnetic field strength. Interestingly, the same coincidence was reported for some of the jovian streams registered by Ulysses¹. On the basis of these findings, we suggest that traversals through CIRs are key to forming the bursts.

Knowledge of the particle mass would allow us to determine the starting position of the grain. This could be achieved by backtracing numerically the trajectories of grains of different sizes from the detector to Saturn, provided that both the IMF and the solar-wind properties are known for this period. Unfortunately, these data are only available until up to 30 January 2004. For impacts during this period, we found that only particles larger than 10 nm were likely to hit the detector (S.K. and S. Hsu, manuscript in preparation). This implies that these particles originated from the outskirts of Saturn's A ring. An interesting consequence of this result is that material from the main rings can be analysed by *in situ* dust detectors like the CDA.

Long-term monitoring of dust streams at Jupiter allowed the separation of the regular behaviour of the system (governed by Jupiter's rotation and Io's orbital motion) from its stochastic variations (due to Io's volcanic activity and the variability of the solar-wind plasma and IMF parameters). Similarly, we expect that monitoring dust streams from Saturn's magnetosphere will be useful in identifying their source, and in understanding both the acceleration mechanism and the coupling between dust (including

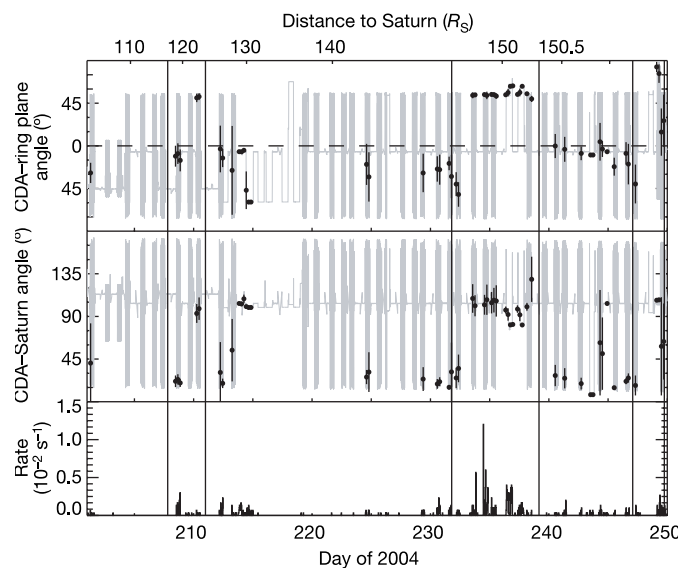


Figure 4 Directionality of particles detected between 18 July (day 200) and 6 September (day 250). Top, the mean angle between CDA and Saturn's ring plane; middle, the mean angle between CDA and Saturn. The mean angle is given whenever more than two impacts within 4 hours were registered. Bottom, the impact rate of CDA. The grey solid lines mark the evolution of the CDA boresight. Throughout the period shown, Cassini was almost continuously performing slow rolls, which are favourable for scanning a large area for dust sources. All impacts detected after SOI are characterised by impact charges too faint to trigger the event recording. For those impacts, the measurement was started by the multiplier signal of the integrated time-of-flight mass spectrometer. This is only possible if the particles hit the small inner target with a field-of-view (FOV) of $\pm 28^\circ$, which is significantly smaller than the total instrument FOV of $\pm 45^\circ$. The vertical error bars represent the standard deviations. The vertical solid lines mark Cassini's traversals through co-rotating interaction regions in the solar wind (M. Dougherty, personal communication), which coincide with the dust bursts.

the rings) and the magnetic field and the plasma environment in Saturn's magnetosphere. □

Received 25 July; accepted 18 November 2004; doi:10.1038/nature03218.

1. Grün, E. *et al.* Discovery of jovian dust streams and interstellar grains by the Ulysses spacecraft. *Nature* **362**, 428–430 (1993).
2. Grün, E. *et al.* Constraints from Galileo observations on the origin of jovian dust streams. *Nature* **381**, 395–398 (1996).
3. Graps, A. L. *et al.* Io as a source of the jovian dust streams. *Nature* **405**, 48–50 (2000).
4. Horányi, M., Morfill, G. & Grün, E. Mechanism for the acceleration and ejection of dust grains from Jupiter's magnetosphere. *Nature* **363**, 144–146 (1993).
5. Hamilton, D. & Burns, J. Ejection of dust from Jupiter's gossamer ring. *Nature* **364**, 695–699 (1993).
6. Horányi, M. Dust stream from Jupiter and Saturn. *Phys. Plasmas* **7**, 3847–3850 (2000).
7. Srama, R. *et al.* The Cassini cosmic dust detector. *Space Sci. Rev.* (in the press).
8. Wehry, A. & Mann, I. Identification of β -meteoroids from measurements of the dust detector onboard the Ulysses spacecraft. *Astron. Astrophys.* **341**, 296–303 (1999).
9. Göller, J. R. & Grün, E. Calibration of the Galileo/Ulysses dust detectors with different projectile materials and at varying impact angles. *Planet. Space Sci.* **37**, 1197–1206 (1989).
10. Zook, H. *et al.* Solar wind magnetic field bending of Jovian dust trajectories. *Science* **274**, 1501–1503 (1996).
11. Horányi, M., Morfill, G. & Grün, E. The dusty ballerina skirt of Jupiter. *J. Geophys. Res.* **98**, 21245–21251 (1993).
12. Richardson, J. D. An extended plasma model for Saturn. *Geophys. Res. Lett.* **22**, 1177–1180 (1995).
13. Chow, V., Mendis, D. & Rosenberg, M. Role of grain size and particle velocity distribution in secondary electron emission in space plasmas. *J. Geophys. Res.* **98**, 19065–19076 (1993).

Acknowledgements This project is supported by the Max-Planck-Institut für Kernphysik and the DLR.

Competing interests statement The authors declare that they have no competing financial interests.

Correspondence and requests for material should be addressed to S.K. (Sascha.Kempf@mpi-hd.mpg.de).

An all-silicon Raman laser

Haisheng Rong¹, Ansheng Liu¹, Richard Jones¹, Oded Cohen², Dani Hak², Remus Nicolaescu¹, Alexander Fang¹ & Mario Paniccia¹

¹Intel Corporation, 2200 Mission College Blvd, CHP3-109, Santa Clara, California 95054, USA

²Intel Corporation, S.B.I. Park Har Hotzvim, Jerusalem, 91031, Israel

The possibility of light generation and/or amplification in silicon has attracted a great deal of attention¹ for silicon-based optoelectronic applications owing to the potential for forming inexpensive, monolithic integrated optical components. Because of its indirect bandgap, bulk silicon shows very inefficient band-to-band radiative electron-hole recombination. Light emission in silicon has thus focused on the use of silicon engineered materials such as nanocrystals^{2–3}, Si/SiO₂ superlattices⁶, erbium-doped silicon-rich oxides^{7–10}, surface-textured bulk silicon¹¹ and Si/SiGe quantum cascade structures¹². Stimulated Raman scattering (SRS) has recently been demonstrated as a mechanism to generate optical gain in planar silicon waveguide structures^{13–21}. In fact, net optical gain in the range 2–11 dB due to SRS has been reported in centimetre-sized silicon waveguides using pulsed pumping^{18–21}. Recently, a lasing experiment involving silicon as the gain medium by way of SRS was reported, where the ring laser cavity was formed by an 8-m-long optical fibre²². Here we report the experimental demonstration of Raman lasing in a compact, all-silicon, waveguide cavity on a single silicon chip. This demonstration represents an important step towards producing practical continuous-wave optical amplifiers and lasers that could be integrated with other optoelectronic components onto CMOS-compatible silicon chips.

The silicon Raman laser described here is based on a low-loss single-mode rib waveguide containing a reverse-biased p-i-n diode structure. The silicon rib waveguide is fabricated on the (100) surface of an undoped silicon-on-insulator (SOI) substrate. Silicon waveguide patterning was done using standard projection photolithography and Cl₂/Ar-based plasma reactive ion etching. A

diagram of the waveguide cross-section is shown in Fig. 1a. The rib waveguide width (W), rib height (H) and etch depth (h) are (in μm) 1.5, 1.55 and 0.7, respectively. The effective core area²³ of the waveguide is calculated to be $\sim 1.6 \mu\text{m}^2$ by using a fully vectorial waveguide modal solver FIMMWAVE (details available at <http://www.photond.com/products/fimmwave.htm>). The waveguide is formed in an S-shaped curve with a bend radius of 400 μm and a total length of 4.8 cm (Fig. 1b). The straight sections of the waveguide are oriented along the [011] direction. The linear optical transmission loss of the S-bend waveguide was measured to be $0.35 \pm 0.05 \text{ dB cm}^{-1}$ on the basis of Fabry-Pérot resonance measurements²⁴ of a polished but uncoated waveguide.

The silicon Raman laser optical cavity was formed by coating one of the waveguide facets with a multi-layer coating while leaving the other facet uncoated. The facet coating was designed to be broadband and have a high reflectivity ($\sim 90\%$) for both the pump wavelength of 1.536 μm and the Raman/Stokes wavelength of 1.67 μm . The uncoated facet has a reflectivity of $\sim 30\%$ for both pump and Raman wavelengths. The pump beam is coupled into the cavity through the uncoated facet (see Fig. 1b). Although the laser beam is emitted from both facets of the silicon chip, we measured the laser output only from the uncoated facet. The waveguide loss was measured before the application of the multi-layer coating to the waveguide facet. The coating reflectivity was determined by measuring the linear optical transmission spectrum of the same waveguides before and after application of the coating.

Figure 1b shows the experimental set-up. An external cavity continuous-wave (c.w.) diode laser at 1.536 μm is amplitude modulated using an acousto-optic modulator to produce square-wave pulses of $\sim 130 \text{ ns}$ width at 10 kHz. An erbium-doped fibre amplifier system of three gain stages is used to amplify these pulses to produce a pump beam of peak power up to 2 W. The pump beam passes through a thin-film-based wavelength de-multiplexer and is coupled into the waveguide cavity by a lensed fibre. The Raman laser output and the reflected pump beam are coupled back into the lensed fibre, and separated through the wavelength de-multiplexer. The extracted laser output from the reflection port of the de-multiplexer is further filtered by a double-grating monochromator before being detected by a power meter or photodetector. The coupling loss between the lensed fibre and the waveguide was measured to be $\sim 4 \text{ dB}$, the insertion loss of the de-multiplexer is 0.6 dB, and the monochromator has a loss of 6.7 dB.

One of the main limiting factors for achieving net gain, and hence lasing, from SRS in silicon waveguides is two-photon generated free carrier absorption (FCA). As has been shown previously^{15,17}, under

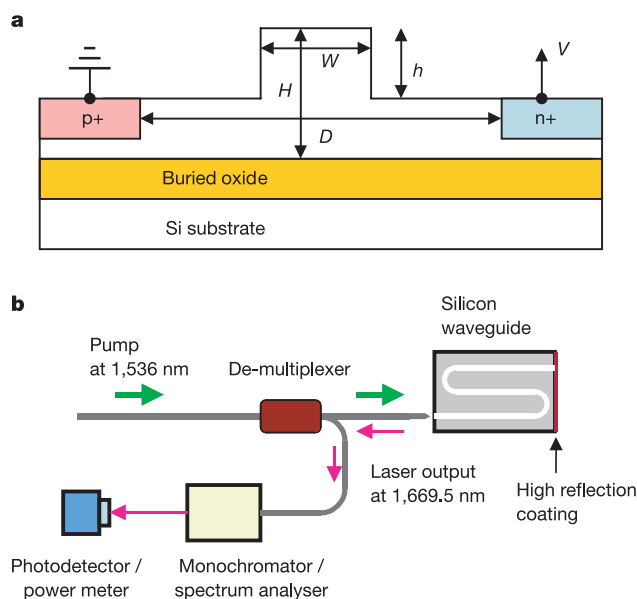


Figure 1 Diagrams showing silicon waveguide laser cavity and experimental set-up. **a**, Cross-section of a silicon-on-insulator (SOI) rib waveguide containing a reverse biased p-i-n diode structure with an applied voltage of V . **b**, Set-up for Raman lasing experiment. See main text for details.

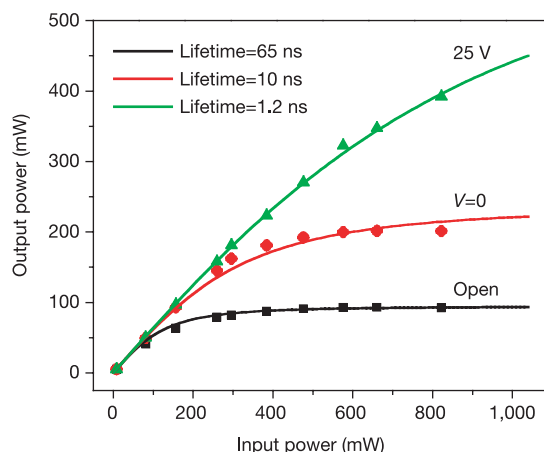


Figure 2 Nonlinear optical transmission of a 4.8-cm-long silicon waveguide containing a p-i-n diode with and without reverse bias. The symbols represent the experimental results and the curves the modelled results. 'Lifetime' indicates effective carrier lifetime.

high-power excitation two photon absorption (TPA) in silicon^{25,26} induces a significant amount of free carriers because of the relatively long carrier recombination lifetime. These photogenerated free carriers induce additional optical loss owing to the free carrier plasma dispersion effect, otherwise known as FCA²⁷. This FCA is problematic, as it both attenuates the pump beam and increases the optical loss for the Raman signal. To reduce such an effect, we designed a p-i-n diode structure along the rib waveguide. The silicon rib waveguide has heavily doped p- and n-type regions in the silicon slab (Fig. 1a). The separation between the p- and n-doped regions is $D = 6 \mu\text{m}$. The p and the n sections of the diode were formed by boron and phosphorus implantation. The implantation dose, energy and thermal treatment resulted in a surface doping concentration of $\sim 1 \times 10^{20} \text{ cm}^{-3}$ as well as small lateral diffusion to minimize interaction with the optical mode. Aluminium films were deposited on the p- and n-doped surface regions to form ohmic contacts (not shown in Fig. 1a) followed by a SiO_2 passivation layer. Our modelling and testing confirm that the linear optical loss of the waveguide is not affected by the heavily doped regions and metal contacts.

When a reverse bias is applied to the p-i-n diode, the TPA-generated electron-hole pairs can be swept out of the silicon waveguide between the p- and n-doped regions by the applied electric field. Thus the carrier transit time or effective carrier lifetime can be modified by the reverse bias voltage. To verify this experimentally, we measured the transmitted power of an anti-reflection-coated silicon p-i-n diode waveguide as a function of the c.w. input power with various reverse bias voltages. The experimental results are shown in Fig. 2. The symbols represent the measured data, and the solid curves are the modelled results based on the formalism described in ref. 17. There is a good agreement between modelled and measured results. As can be seen, the transmitted power for a silicon waveguide saturates to a given level as the input power is increased. This is a result of TPA-induced FCA. Increasing the bias voltage on the p-i-n diode allows the transmission of the device to be improved. In the modelling, the carrier lifetime is used as a fitting parameter. The linear optical loss is $\sim 0.35 \text{ dB cm}^{-1}$, the TPA coefficient is 0.5 cm GW^{-1} (ref. 14) and the FCA cross-section is $1.45 \times 10^{-17} \text{ cm}^2$ at the wavelength of $1.55 \mu\text{m}$ (refs 24, 27). The anode (p-type region) of the p-i-n is always connected to ground. When the p-i-n diode circuit is open (that is, the p- and n-doped regions are not wire connected so that no net current is flowing), significant saturation in transmission is observed, and the modelled carrier lifetime is 65 ns. When the p-i-n diode circuit is closed (current flow is allowed but the bias voltage is zero), transmission is clearly improved, and the modelled carrier lifetime is 10 ns. When the p-i-n diode is reverse biased with a voltage of 25 V, the power

transmission is further improved, and the corresponding carrier lifetime becomes 1.2 ns. These results show that the nonlinear loss due to the TPA-induced FCA is much reduced by shortening the carrier lifetime using the reverse biased p-i-n diode structure.

The single-pass Raman gain is measured for a silicon p-i-n waveguide with anti-reflection coatings on both facets using the techniques described in refs 17 and 18. With a reverse bias of 25 V applied to the p-i-n waveguide, a single-pass peak on-off gain of 5.2 dB is obtained, which is sufficient to achieve lasing in the cavity described above with a total round-trip loss of 9 dB. From these gain measurements, the Raman gain coefficient is estimated to be $\sim 7.5 \text{ cm GW}^{-1}$ using the method given in ref. 18.

Figure 3 shows the average power of the laser output measured from the uncoated side of the silicon waveguide cavity with a reverse bias of 25 V as a function of average pump power into the waveguide cavity, as depicted in Fig. 1b. As can be seen from Fig. 3, the laser threshold is at $\sim 0.4 \text{ mW}$, and the slope efficiency (single side output) is 9.4%. Including the laser output from the other facet with 90% reflectivity, we estimate the total slope efficiency to be $\sim 10\%$. At a pump power $> 0.9 \text{ mW}$, the laser output starts to saturate. This mainly results from the fact that the free carrier lifetime is of the order of $\sim 1 \text{ ns}$ and nonlinear loss due to the TPA-induced FCA reduces the net gain at high pump powers.

At a pump power of 0.7 mW , we measured the spectrum of the Raman laser from the waveguide cavity. The measurement results are shown in Fig. 4a. For comparison we also show the spectrum of the spontaneous Raman emission taken on an identical waveguide but without the highly reflective coatings. We see from Fig. 4a that the laser spectral width is much narrower than that of the spontaneous emission spectrum. Figure 4b is a higher-resolution

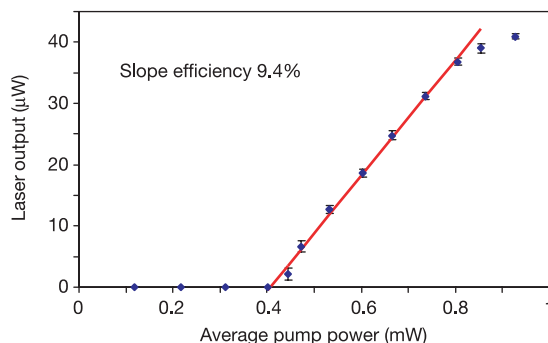


Figure 3 Average output of the silicon Raman laser with reverse biased p-i-n diode as a function of the average pump power into the silicon waveguide cavity. The input pump wavelength is $1,536 \text{ nm}$ and the output Raman laser wavelength is $1,669.5 \text{ nm}$. The reverse bias voltage is 25 V. The slope efficiency (single side output) is 9.4%. Error bars represent standard deviations.

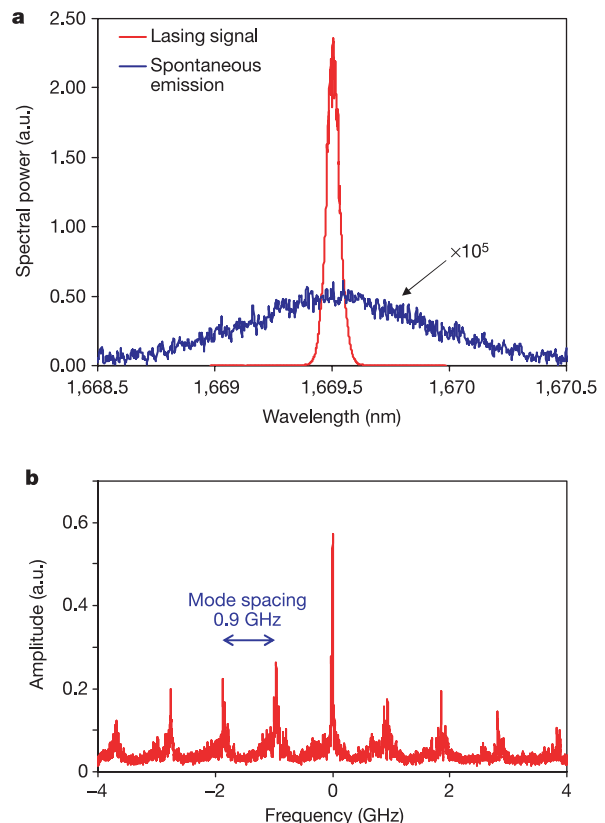


Figure 4 Silicon Raman laser spectra. **a**, Laser spectrum at an average pump power of 0.7 mW (red trace). For comparison, the blue trace shows the spontaneous Raman emission spectrum from a waveguide without a cavity. Note the different vertical scales. **b**, High-resolution Raman laser spectrum measured with a scanning Fabry-Pérot interferometer, showing resolved cavity modes.

spectrum of the laser using a scanning Fabry–Pérot spectrum analyser. We see that the laser has multiple cavity modes with a mode spacing of ~ 0.9 GHz, which corresponds to the free spectral range of a 4.8-cm-long silicon waveguide cavity.

This demonstration of Raman lasing in a silicon waveguide cavity on a single chip represents a significant step towards a more practical, all-silicon-based, c.w. amplifier or laser. We note that the multi-layer coatings are not essential for this device, and that single chip resonators may be fabricated using waveguide Bragg reflectors or ring resonator architectures; both these alternatives are compatible with CMOS processing. Further reduction in the carrier lifetime or cavity optimization could lead to c.w. Raman lasing in silicon. □

Received 23 October; accepted 10 December 2004; doi:10.1038/nature03273.
Published online 5 January 2005.

1. Pavesi, L., Gaponenko, S. & Dal Negro, L. (eds) *Towards the First Silicon Laser* (NATO science series, Kluwer, Dordrecht, 2003).
2. Shimizu-Iwayama, T. *et al.* Visible photoluminescence in Si³⁺-implanted silica glass. *J. Appl. Phys.* **75**, 7779–7783 (1994).
3. Brongersma, M. L., Polman, A., Min, K. S., Tambo, T. & Atwater, H. A. Tuning the emission wavelength of Si nanocrystals in SiO₂ by oxidation. *Appl. Phys. Lett.* **72**, 2577–2579 (1998).
4. Iacona, F., Franzo, G. & Spinella, C. Correlation between luminescence and structural properties of Si nanocrystals. *J. Appl. Phys.* **87**, 1295–1303 (2000).
5. Pavesi, L., Negro, L. D., Mazzoleni, C., Franzo, G. & Priolo, F. Optical gain in silicon nanocrystals. *Nature* **408**, 440–444 (2000).
6. Lockwood, D. J., Lu, Z. H. & Baribeau, J. M. Quantum confined luminescence in Si/SiO₂ superlattices. *Phys. Rev. Lett.* **76**, 539–541 (1996).
7. Lombardo, S., Campisano, S. U., van den Hoven, G. N., Cacciato, A. & Polman, A. Room-temperature luminescence from Er³⁺-implanted semi-insulating polycrystalline silicon. *Appl. Phys. Lett.* **63**, 1942–1944 (1993).
8. Fujii, M., Yoshida, M., Kanzawa, Y., Hayashi, S. & Yamamoto, K. 1.54 μ m photoluminescence of Er³⁺-doped into SiO₂ films containing Si nanocrystals: evidence for energy transfer from Si nanocrystals to Er³⁺. *Appl. Phys. Lett.* **71**, 1198–1200 (1997).
9. Kik, P. G., Brongersma, M. L. & Polman, A. Strong exciton-erbium coupling in Si nanocrystal-doped SiO₂. *Appl. Phys. Lett.* **76**, 2325–2327 (2000).
10. Han, H. S., Seo, S. Y. & Shin, J. H. Optical gain at 1.54 μ m in erbium-doped nanocluster sensitized waveguide. *Appl. Phys. Lett.* **79**, 4568–4570 (2001).
11. Trupke, T., Zhao, J., Wang, A., Corkish, R. & Green, M. Very efficient light emission from bulk crystalline silicon. *Appl. Phys. Lett.* **82**, 2996–2998 (2003).
12. Dehlinger, G. *et al.* Intersubband electroluminescence from silicon-based quantum cascade structures. *Science* **290**, 2277–2280 (2000).
13. Claps, R., Dimitropoulos, D., Han, Y. & Jalali, B. Observation of Raman emission in silicon waveguides at 1.54 μ m. *Opt. Express* **10**, 1305–1313 (2002).
14. Claps, R., Dimitropoulos, D., Raghunathan, V., Han, Y. & Jalali, B. Observation of stimulated Raman amplification in silicon waveguides. *Opt. Express* **11**, 1731–1739 (2003).
15. Liang, T. K. & Tsang, H. K. Role of free carriers from two-photon absorption in Raman amplification in silicon-on-insulator waveguides. *Appl. Phys. Lett.* **84**, 2745–2747 (2004).
16. Espinola, R. L., Dadap, J. L., Osgood, R. M., McNab, S. J. & Vlasov, Y. A. Raman amplification in ultrasmall silicon-on-insulator wire waveguides. *Opt. Express* **12**, 3713–3718 (2004).
17. Rong, H. *et al.* Raman gain and nonlinear optical absorption measurement in a low loss silicon waveguide. *Appl. Phys. Lett.* **85**, 2196–2198 (2004).
18. Liu, A., Rong, H., Panickia, M., Cohen, O. & Hak, D. Net optical gain in a low loss silicon-on-insulator waveguide by stimulated Raman scattering. *Opt. Express* **12**, 4261–4267 (2004).
19. Xu, Q., Almeida, V. & Lipson, M. Time-resolved study of Raman gain in highly confined silicon-on-insulator waveguides. *Opt. Express* **12**, 4437–4442 (2004).
20. Liang, T. K. & Tsang, H. K. Efficient Raman amplification in silicon-on-insulator waveguides. *Appl. Phys. Lett.* **85**, 3343–3345 (2004).
21. Boyraz, O. & Jalali, B. Demonstration of 11dB fiber-to-fiber gain in a silicon Raman amplifier. *IEEE Elect. Express* **1**, 429–434 (2004).
22. Boyraz, O. & Jalali, B. Demonstration of a silicon Raman laser. *Opt. Express* **12**, 5269–5273 (2004).
23. Agrawal, G. P. *Nonlinear Fiber Optics* 2nd edn (Academic, New York, 1995).
24. Reed, G. T. & Knights, A. P. *Silicon Photonics: An Introduction* (John Wiley, Chichester, UK, 2004).
25. Tsang, H. K. *et al.* Optical dispersion, two photon absorption and self-phase modulation in silicon waveguides at 1.5 μ m wavelength. *Appl. Phys. Lett.* **80**, 416–418 (2002).
26. Dinu, M., Quochi, F. & Garcia, H. Third-order nonlinearities in silicon telecom wavelengths. *Appl. Phys. Lett.* **82**, 2954–2956 (2003).
27. Soref, R. A. & Lorenzo, P. J. All-silicon active and passive guided-wave components for $\lambda = 1.3$ and 1.6 μ m. *IEEE J. Quant. Electron.* **QE-22**, 873–879 (1986).

Acknowledgements We thank A. Alduino, D. Tran, J. Tseng, D. Hodge and J. Johnson for assistance in device fabrication and sample preparation; S. Koehl for software development; M. Morse, H. Liu, M. Salib, D. Samarubio, L. Liao, R. Li and G. Ding for technical discussions; and G. T. Reed, I. P. Kaminow and J. E. Bowers for conversations.

Competing interests statement The authors declare that they have no competing financial interests.

Correspondence and requests for materials should be addressed to H.R. (haisheng.rong@intel.com).

Stable sea surface temperatures in the western Pacific warm pool over the past 1.75 million years

Thibault de Garidel-Thoron¹, Yair Rosenthal^{1,2}, Franck Bassinot³ & Luc Beaufort⁴

¹Institute of Marine and Coastal Sciences-Rutgers University, 71 Dudley Road, and

²Department of Geological Sciences-Rutgers University, New Brunswick, New Jersey 08901, USA

³Laboratoire des Sciences du Climat et de l'Environnement-CEA, 91198 Gif-sur-Yvette cedex, France

⁴CEREGE-CNRS Université Aix-Marseille 3, BP80, 135345 Aix-en-Provence cedex 4, France

About 850,000 years ago, the period of the glacial cycles changed from 41,000 to 100,000 years. This mid-Pleistocene climate transition has been attributed to global cooling, possibly caused by a decrease in atmospheric carbon dioxide concentrations^{1,2}. However, evidence for such cooling is currently restricted to the cool upwelling regions in the eastern equatorial oceans^{3,4}, although the tropical warm pools on the western side of the ocean basins are particularly sensitive to changes in radiative forcing^{5,6}. Here we present high-resolution records of sea surface temperatures spanning the past 1.75 million years, obtained from oxygen isotopes and Mg/Ca ratios in planktonic foraminifera from the western Pacific warm pool. In contrast with the eastern equatorial regions, sea surface temperatures in the western Pacific warm pool are relatively stable throughout the Pleistocene epoch, implying little long-term change in the tropical net radiation budget. Our results challenge the hypothesis of a gradual decrease in atmospheric carbon dioxide concentrations as a dominant trigger of the longer glacial cycles since 850,000 years ago. Instead, we infer that the temperature contrast across the equatorial Pacific Ocean increased, which might have had a significant influence on the mid-Pleistocene climate transition.

The warmest pool of oceanic surface lies in the western equatorial Pacific Ocean and plays an important role in Earth's climate. Annual sea surface temperatures (SSTs) above 28 °C (ref. 7) lead to deep atmospheric convection, which transfers through the atmosphere large amounts of water vapour and latent heat from tropical to mid-latitudes. The importance of the western equatorial Pacific warm pool (WPWP) is perhaps best manifested by the impact of the El Niño/Southern Oscillation (ENSO) on the interannual variability in the Earth's temperature and precipitation patterns^{8,9}. ENSO-linked changes in tropical SST patterns also have a major influence on heat transport from low to mid- and high latitudes¹⁰ and models suggest that long-term changes of SST patterns in the equatorial Pacific may also have had large consequences for extra-tropical climate in the past¹¹. Thus, reconstructing past changes in tropical Pacific SSTs is critical for understanding the evolution of global heat and moisture transport, such as those occurring during the mid-Pleistocene transition, characterized by the onset of the large 100-kyr glacial cycles, about 0.85 Myr ago.

Here, we present a high-resolution Pleistocene record (0–1.75 Myr ago, Marine Isotopic Stages MIS1 to MIS59) of paired $\delta^{18}\text{O}$ and Mg/Ca measurements in planktonic foraminifera (*Globigerinoides ruber*, 250–300 μ m, white) (Fig. 1) from the IMAGES core MD97-2140 located in the heart of the WPWP (2° 02' N, 141° 46' E, 2,547 m; Supplementary Fig. S1). The record extends the previously published, late-Pleistocene Ocean Drilling Program (ODP) site 806B record (0–450 kyr ago) from the Ontong-Java plateau¹², thus providing new insights on tropical Pacific climate variability during

the early Pleistocene. The records demonstrate that the mid-Pleistocene transition from 41-kyr- to 100-kyr-dominated periodicities, previously shown in high-latitude cores and more recently in the eastern equatorial Pacific⁴, is also a prominent feature of the WPWP climate. Spectral analysis of the $\delta^{18}\text{O}$ and Mg/Ca records shows a dominant 41-kyr peak between 1,478 and 900 kyr ago and a 100-kyr peak between 850 and 6 kyr ago (Fig. 2). Thus, the WPWP record indicates that the early-Pleistocene obliquity-related oscillation observed in the eastern equatorial Pacific⁴ was not limited to upwelling regions, but rather included the entire tropical Pacific. The onset of the 100-kyr cycle in the WPWP is accompanied by a significant increase in the amplitude of the glacial–interglacial $\delta^{18}\text{O}_{G.ruber}$. Between 1.75 and 0.85 Myr ago the glacial–interglacial amplitude (see Methods) in MD97-2140 is about $0.6 \pm 0.2\text{‰}$, whereas during the late Pleistocene the amplitude increased to about $1 \pm 0.3\text{‰}$, reaching 1.4‰ during terminations I, II and IV. Although the shift in the dominant frequency and glacial–interglacial amplitude at about 0.85 Myr ago is consistent with other Pleistocene records (see ref. 13 for example), the WPWP record is unique in that it shows no discernible long-term trend in $\delta^{18}\text{O}_{G.ruber}$ and remains at about -1.8‰ with a variance around this mean of $\pm 0.6\text{‰}$ throughout the Pleistocene.

The Mg/Ca-derived SST record also exhibits the change in periodicity 0.85 Myr ago (Fig. 1b) and, like $\delta^{18}\text{O}$, shows no long-term trend throughout the Pleistocene. As the core is situated above the modern lysocline, dissolution effects on our SST record are considered to be minimal. In support of this conclusion, our late-Holocene estimate of $\sim 28.7 \pm 0.5^\circ\text{C}$ is in close agreement with the modern mean annual SST of 28.6°C in this region⁷, and our

estimated $\sim 2.7 \pm 0.5^\circ\text{C}$ cooling of the tropical Pacific during the last and penultimate glaciations accord with previous studies^{12,14}. On a longer timescale, the late-Pleistocene part of our record is generally consistent with the ODP806B SST record from the Ontong-Java plateau¹². The amplitude of a full glacial SST cycle in the WPWP is on average $2.2 \pm 0.7^\circ\text{C}$ during the late Pleistocene (6–850 kyr ago), and $1.5 \pm 0.6^\circ\text{C}$ during the early Pleistocene (850–1,750 kyr ago). As glacial SSTs are relatively constant at $25.9 \pm 0.2^\circ\text{C}$ throughout the Pleistocene, the changes in glacial–interglacial amplitude reflect mainly variability in interglacial maxima. Throughout the record, the increase in SST during glacial terminations leads the deglacial change in $\delta^{18}\text{O}_{G.ruber}$. Cross-spectral analysis shows that between 0 and 0.85 Myr ago the SST lead is about 4 ± 4 kyr in the eccentricity band, in agreement with results from other late-Pleistocene equatorial Pacific records^{4,12,14}. Between 0.85 and 1.75 Myr ago the SST lead is about 10 ± 4 kyr in the dominant 41-kyr band (Fig. 2). The early-Pleistocene average SST is $27.1 \pm 1^\circ\text{C}$, which is not statistically different from the late-Pleistocene average SST of $27.2 \pm 1^\circ\text{C}$; and indicates that long-term temperature in the WPWP is not influenced by the Mid-Pleistocene Revolution.

Our record of surface water $\delta^{18}\text{O}_{\text{water}}$ (used as a proxy of surface salinity; Fig. 1), calculated from Mg/Ca and $\delta^{18}\text{O}_{G.ruber}$ data, provides a useful assessment of hydrological variations in the WPWP across the mid-Pleistocene transition. The record indicates the long-term stability of $\delta^{18}\text{O}_{\text{water}}$ (average $0.66 \pm 0.25\text{‰}$ and $0.65 \pm 0.25\text{‰}$ for early and late Pleistocene, respectively). This contrasts with the record of deep-water benthic foraminiferal $\delta^{18}\text{O}$ (see, for example, Fig. 3), which shows a long-term trend of

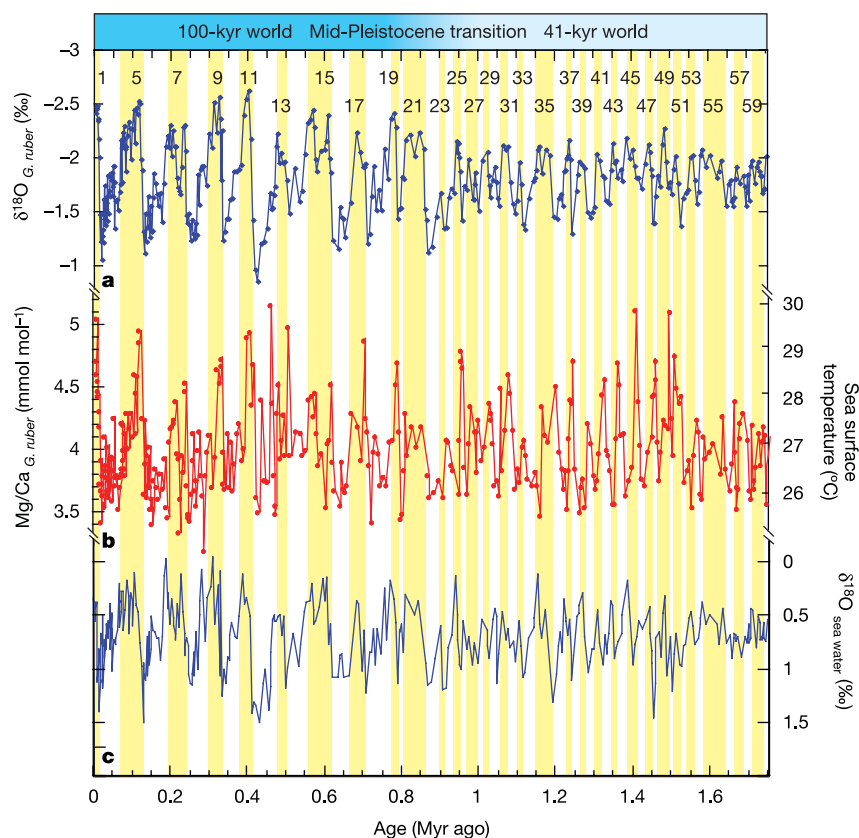


Figure 1 Pleistocene reconstruction of surface water changes in the WPWP from the IMAGES core MD97-2140. **a**, **b**, Planktonic foraminiferal records (*G. ruber*) of $\delta^{18}\text{O}$ (**a**) and Mg/Ca SST (**b**) from IMAGES core MD97-2140 (Eauripik rise $2^\circ 02' \text{N}$, $141^\circ 46' \text{E}$, 2,547 m). SST estimates were calculated using the core-top calibration²⁹ of Mg/Ca = $0.30 \times \exp(0.095 \times \text{SST})$. **c**, $\delta^{18}\text{O}_{\text{sea water}}$ computed using the isotopic equation of

ref. 30. Interglacial periods are numbered following the terminology of SPECMAP (ref. 13 and references therein). Note the concomitant increase in interglacial $\delta^{18}\text{O}_{G.ruber}$ with decreasing SSTs between 1.2 and 0.85 Myr ago, before the onset of the late Pleistocene 100-kyr glaciations.

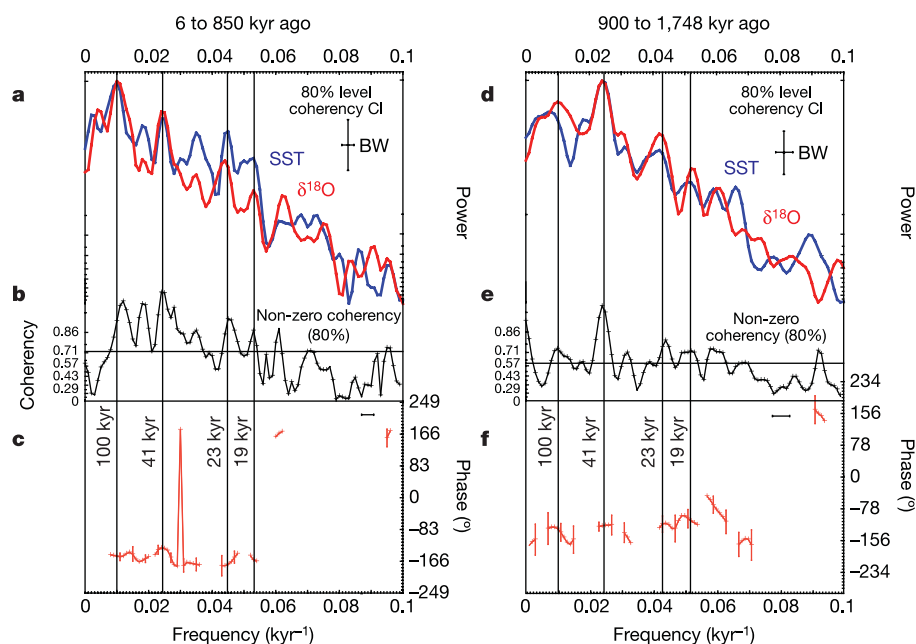


Figure 2 Comparison of spectral analysis of $\delta^{18}\text{O}$ and Mg/Ca-derived SST records from MD97-2140 in the late and early Pleistocene. **a**, Time interval from 6 to 850 kyr ago; **b**, time interval from 900 to 1,748 kyr ago. Panels also present the coherence (**b**, **e**) and

phase relationship (**c**, **f**) obtained from cross-spectral analysis between the MD97-2140 $\delta^{18}\text{O}$ and SST for each period. The confidence interval (CI) at the 80% significance level and the bandwidth (BW) used for the spectral analysis are displayed.

increasing $\delta^{18}\text{O}$ through the Pleistocene, presumably reflecting expansion of continental ice. We suggest that the effect of Pleistocene ice-sheet growths on surface WPWP $\delta^{18}\text{O}$ was compensated by progressive regional freshening. Nowadays, although advection of surface water might also have affected the hydrography of the WPWP, the surface salinity is mainly influenced by the local precipitation–evaporation balance¹⁵. Our data suggest that although the thermal response of the WPWP to the mid-Pleistocene transition changes was small, there was a substantial effect on the regional hydrological cycle.

Recently, it has been suggested that the increase in the Pacific zonal SST contrast and development of strong Walker circulation—driven by enhanced upwelling and cooling of surface waters in the eastern equatorial Pacific (ODP site 851, 2° 46' N, 110° 34' W)—might have had a strong influence on the climate response to radiative changes, thus contributing to the Pliocene cooling trend¹⁶. A further cooling of the eastern equatorial Pacific cold tongue, also attributed to changes in upwelling, occurred during the mid-Pleistocene (between 1.2 and 0.8 Myr ago), as inferred from an alkenone-based SST record of ODP site 846 (3° 6' S, 90° 49' W)⁴. Our new record (MD97-2140) suggests, however, that this cooling was restricted to the upwelling region and did not affect the greater tropical band. This conclusion is supported by evidence from the Atlantic Ocean, where a subtropical record from the Benguela upwelling region shows significant cooling during the mid-Pleistocene transition³ (Fig. 3), whereas another subtropical record of ODP site 1077 (10° 26' E, 5° 10' S) does not show any long-term SST change across the mid-Pleistocene transition¹⁷. Combined, these observations suggest that the mid-Pleistocene cooling was restricted to upwelling regions, whereas surface temperatures remained at the same mean level before and after 0.8 Myr ago in the WPWP and probably in most of the tropical oceans.

In the modern ocean, the interannual variability in the temperature difference (ΔSST) between the sites of MD97-2140 and ODP846 is highly correlated with the Southern Oscillation index ($r^2 \approx 0.6$), an index of the ENSO; a larger ΔSST , primarily a result of enhanced upwelling in the eastern equatorial Pacific, occurs during the cold phase of ENSO (La Niña) and a smaller ΔSST

during the warm phase (El Niño). La Niña events, which are periods of intensified low-latitude zonal winds, and enhanced precipitation in the western equatorial Pacific are also characterized by fresher sea surface waters at the site of the MD97-2140 (ref. 18). Although a direct comparison of the alkenone-derived SST record from ODP846 and the Mg/Ca SST record of MD97-2140 might be problematic owing to possible discrepancies between the two proxies¹⁹, recent multi-proxy studies indicate that their respective estimates generally agree to within 1 °C in the equatorial Pacific²⁰. Taken at face value, the comparison indicates a long-term change in the ΔSST between the cold tongue and the warm pool from the early to late Pleistocene (Fig. 3). The most notable feature of the ΔSST reconstruction is an increase from ~3.5 to ~5 °C that occurred during the interval between 1.2 and 0.8 Myr ago, coincident with the mid-Pleistocene increase in ice volume as reflected in the positive shift in the benthic foraminiferal $\delta^{18}\text{O}$ record of ODP846 (Fig. 3). The increase in ΔSST gradient between the eastern and western equatorial Pacific sites and inferred freshening at MD97-2140 are consistent with a change in the mean climate state from an El Niño-like state to more La Niña-like conditions, which implies an intensification of the Walker circulation during the mid-Pleistocene (Fig. 3).

Most mechanisms proposed to explain the mid-Pleistocene shift from 41-kyr to 100-kyr periodicity in climate records invoke a gradual global cooling during the Pleistocene, until a threshold is passed around 0.85 Myr ago, when feedbacks exerted by continental and/or marine ice-sheet dynamics are assumed to have become preponderant^{1,21,22}. This gradual global cooling scenario is attributed by some to a decrease in atmospheric p_{CO_2} , and its effect on radiative forcing, during the early Pleistocene¹. Evidence for this 'cooling' hypothesis comes primarily from alkenone-based SST reconstructions from the upwelling regions, but is not supported by our WPWP record. Moreover, as the long-term control of equatorial Pacific SSTs results from the combined effects of greenhouse gases (mainly water vapour, CO_2 and CH_4) and albedo^{5,6}, our new SST record provides a different perspective on Pleistocene climate evolution. The long-term stability of WPWP temperatures during the Pleistocene implies little change in the tropical net

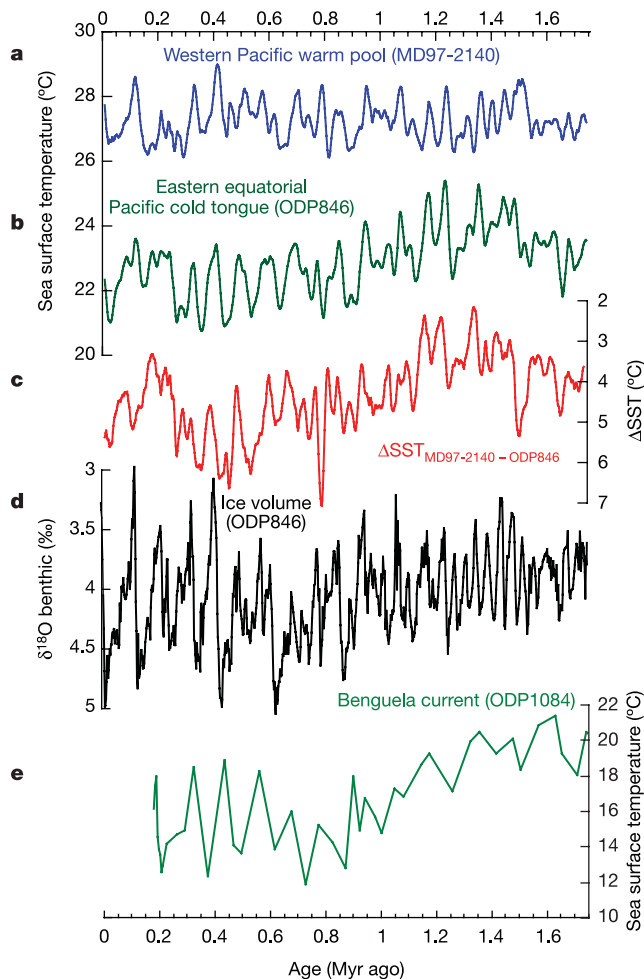


Figure 3 Comparison of the WPWP Pleistocene climatic evolution with palaeoclimatic records from different regions. **a**, **b**, Mg/Ca-derived SST record from MD97-2140 (**a**) in the WPWP and the alkenone-derived SST record of ODP846 (**b**) from the cold tongue in the eastern equatorial Pacific⁴. Both records are smoothed with a 20-kyr moving average; **c**, shows the $\Delta\text{SST}_{\text{MD97-2140}-\text{ODP846}}$ record as a proxy of the zonal surface thermal gradient between the western and the eastern equatorial Pacific. Note the reversed temperature scale. **d**, The eastern equatorial Pacific $\delta^{18}\text{O}$ benthic (site ODP 846) is used as a proxy of ice volume²⁸. **e**, Smoothed with a 20-kyr moving average, the Benguela current upwelling alkenone-based SST record³.

radiative budget. This relative stability challenges the decreasing atmospheric p_{CO_2} scenario assumed for the onset of the mid-Pleistocene large 100-kyr glaciations¹.

An alternative view is that nonlinear coupled ocean-atmosphere dynamics that regulate the equatorial Pacific climate, might have played a more significant role than changes in radiative forcing in the mid-Pleistocene transition; in agreement with recent studies of the Pliocene initiation of Northern Hemisphere glaciation¹⁶. The increased temperature contrast between the eastern equatorial Pacific upwelling regions and the thermally stable WPWP in the mid-Pleistocene probably led to the intensification of the Walker circulation. This strengthening of the zonal Pacific atmospheric circulation is further supported by the surface freshening of the WPWP. Both observations imply drastic changes in the equatorial Pacific SST and salinity patterns across the mid-Pleistocene transition. These changes might have led to atmospheric processes¹¹ and oceanic interactions²³, which could have altered the meridional heat and moisture transfer to the Northern Hemisphere ice sheets during the mid-Pleistocene transition and thus contributed to the Northern Hemisphere glaciation. □

Methods

Analytical methods

Isotope measurements were done at LSCE (Gif-sur-Yvette, France), using a Kiel device coupled to a Finnigan MAT 251 mass spectrometer. The external precision of $\delta^{18}\text{O}$ analysis was 0.05‰, as determined by repeated measurements of a standard carbonate material, NBS19. Mg/Ca measurements were done at Rutgers University using Finnigan Element 1 inductively coupled plasma mass spectrometry (ICP-MS), following the analytical protocol detailed in ref. 24. Samples were prepared using the Boyle and Keigwin method²⁵, excluding the reductive step. The long-term external precision of Mg/Ca analysis was about 4.5% (1 σ , relative standard deviation, r.s.d.) as determined by repeated measurements of three consistency standards with Mg/Ca ratios varying from 1.2 to 7 mmol mol⁻¹. This is equivalent to $\pm 0.5^\circ\text{C}$ analytical error in the core record. The planktonic foraminifer *G. ruber* was chosen for this study because this species lives strictly in the mixed layer, and therefore its geochemical composition reflects the overlying surface hydrography. Moreover, fluxes of *G. ruber* to the sediment in the WPWP do not show any evidence of a seasonal bias, circumventing the possibility of a seasonal effect on the estimated SST²⁶.

Age model and time-series analyses

The MD97-2140 $\delta^{18}\text{O}_{\text{G. ruber}}$ record was tuned to the astronomically calibrated ODP677 $\delta^{18}\text{O}_{\text{G. ruber}}$ record¹³ located in the eastern equatorial Pacific (Fig. 1), yielding chronology consistent with major micro-paleontological (disappearance of *G. ruber* pink variety) and palaeomagnetic (Brunhes-Matuyama boundary²⁷) events. The average sedimentation rate is $\sim 2\text{ cm kyr}^{-1}$ throughout the last 1,755 kyr recovered by the core. The core was sampled every 10 cm for both isotopic and trace-metal analyses, yielding $\sim 5\text{-kyr}$ resolution. To calculate the equatorial zonal temperature gradient, we first matched the MD97-2140 $\delta^{18}\text{O}_{\text{G. ruber}}$ record to the ODP846 $\delta^{18}\text{O}_{\text{benthic}}$ record²⁸ to get a consistent timescale for both records. The zonal thermal gradient was calculated as the difference: $\Delta\text{SST} = \text{SST}_{\text{MD97-2140}} - \text{SST}_{\text{ODP846}}$. To compute the amplitude of glacial-interglacial changes in SST and $\delta^{18}\text{O}$, we first defined the MIS using the $\delta^{18}\text{O}_{\text{G. ruber}}$ correlation to the ODP846 benthic reference curve²⁸. Then we averaged the downcore values within each MIS. The amplitude of glacial-interglacial variations was finally computed by subtracting the interglacial estimates from the glacial ones. Spectral analyses were performed using the Arand software package of Brown University (available at <http://www.ngdc.noaa.gov/paleo/softlib/>), after linear detrending of time series. All MD97-2140 data are archived on the NGDC-NOAA palaeoclimate database website: <http://www.ngdc.noaa.gov/paleo/paleo.html>.

Received 12 July; accepted 15 November 2004; doi:10.1038/nature03189.

- Berger, A., Li, X. S. & Loutre, M.-F. Modelling northern hemisphere ice volume over the last 3 Ma. *Quat. Sci. Rev.* **18**, 1–11 (1999).
- Raymo, M. E. The timing of major climate terminations. *Paleoceanography* **12**, 577–585 (1997).
- Marlow, J. R., Lange, C. B., Wefer, G. & Rosell-Mele, A. Upwelling intensification as part of the Pliocene-Pleistocene climate transition. *Science* **290**, 2288–2291 (2000).
- Liu, Z. & Herbert, T. D. High-latitude influence on the eastern equatorial Pacific climate in the early Pleistocene epoch. *Nature* **427**, 720–723 (2004).
- Lea, D. W. The 100,000 year cycle in tropical SST, greenhouse forcing, and climate sensitivity. *J. Clim.* **17**, 2170–2179 (2004).
- Broccoli, A. J. Tropical cooling at the last glacial maximum: an atmosphere-mixed layer ocean model simulation. *J. Clim.* **13**, 951–976 (2000).
- Conkright, M. et al. *World Ocean Atlas 1998* [CD-ROM] (Data Set Documentation 16, NODC, Silver Springs, Maryland, 1998).
- Ropelewski, C. F. & Halpert, M. S. Global and regional scale precipitation patterns associated with the El Niño/Southern Oscillation. *Mon. Weath. Rev.* **115**, 1606–1626 (1986).
- Halpert, M. S. & Ropelewski, C. F. Surface temperature patterns associated with the Southern Oscillation. *J. Clim.* **5**, 577–593 (1992).
- Sun, D. Z. Possible effect of an increase in the warm-pool SST on the magnitude of El Niño warming. *J. Clim.* **16**, 185–205 (2003).
- Yin, J. H. & Battisti, D. S. The importance of tropical sea surface temperature patterns in simulations of Last Glacial Maximum climate. *J. Clim.* **14**, 565–581 (2001).
- Lea, D. W., Pak, D. K. & Spero, H. J. Climate impact of late quaternary equatorial Pacific sea surface temperature variations. *Science* **289**, 1719–1724 (2000).
- Shackleton, N. J., Berger, A. & Peltier, W. R. An alternative astronomical calibration of the lower Pleistocene timescale based on ODP Site 677. *Trans. R. Soc. Edinb. Earth Sci.* **81**, 251–261 (1990).
- Visser, K., Thunell, R. C. & Stott, L. Magnitude and timing of temperature change in the Indo-Pacific warm pool during deglaciation. *Nature* **421**, 152–155 (2003).
- Delcroix, T. Observed surface oceanic and atmospheric variability in the tropical Pacific at seasonal and ENSO timescales: a tentative overview. *J. Geophys. Res.* **103**, 18611–18633 (1998).
- Ravelo, A. C., Andressen, D. H., Lyle, M., Lyle, A. O. & Wara, M. W. Regional climate shifts caused by gradual global cooling in the Pliocene epoch. *Nature* **429**, 263–267 (2004).
- Scheffé, E., Schouten, S., Fred Jansen, J. H. & Sinninghe Damste, J. S. African vegetation controlled by tropical sea surface temperatures in the mid-Pleistocene period. *Nature* **422**, 418–421 (2003).
- Delcroix, T. & McPhaden, M. Interannual sea surface salinity and temperature changes in the western Pacific warm pool during 1992–2000. *J. Geophys. Res. Oceans* **107**, 22135–22150 (2002).
- Nürnberg, D., Mueller, A. & Schneider, R. R. Paleo-sea surface temperature calculations in the equatorial east Atlantic from Mg/Ca ratios in planktic foraminifers: A comparison to sea surface temperature estimates from Uk'37, oxygen isotopes, and foraminiferal transfer function. *Paleoceanography* **15**, 124–134 (2000).
- Kiefer, T. & Kienast, M. Patterns of deglacial warming in the Pacific Ocean: a review with emphasis on the time interval of Heinrich event 1. *Quat. Sci. Rev.* (in the press).
- Imbrie, J. et al. On the structure and origin of major glaciation cycles. 2. The 100,000-year cycle. *Paleoceanography* **8**, 699–735 (1993).
- Tziperman, E. & Gildor, H. On the mid-Pleistocene transition to 100-kyr glacial cycles and the

- asymmetry between glaciation and deglaciation times. *Paleoceanography* 18, doi:10.1029/2001PA000627 (2003).
23. Philander, S. G. & Fedorov, A. V. Role of tropics in changing the response to Milankovitch forcing some three million years ago. *Paleoceanography* 18, doi:10.129/2002PA000837 (2003).
24. Rosenthal, Y., Field, F. & Sherrell, R. M. Precise determination of element/calcium ratios in calcareous samples using sector field inductively coupled plasma mass spectrometry. *Anal. Chem.* 71, 3248–3253 (1999).
25. Boyle, E. A. & Keigwin, L. D. Comparison of Atlantic and Pacific paleochemical records for the last 250,000 years: changes in deep ocean circulation and chemical inventories. *Earth Planet. Sci. Lett.* 76, 135–150 (1985).
26. Kawahata, H., Nishimura, A. & Gagan, M. K. Seasonal change in foraminiferal production in the western equatorial Pacific warm pool: evidence from sediment trap experiments. *Deep-sea Res.* 49, 2783–2800 (2002).
27. Carcaillet, J. T., Thouveny, N. & Bourles, D. L. Geomagnetic moment instability between 0.6 and 1.3 Ma from cosmogenic evidence. *Geophys. Res. Lett.* 30, doi:10.1029/2003GL017550 (2003).
28. Mix, A. C. *et al.* Benthic foraminiferal stable isotope stratigraphy of Site 846; 0–1.8 Ma. *Proc. ODP Sci. Res.* 138, 839–854 (1995).
29. Rosenthal, Y. & Lohmann, G. P. Accurate estimation of sea surface temperatures using dissolution-corrected calibrations for Mg/Ca paleothermometry. *Paleoceanography* 17, doi: 1029/2001PA000749 (2002).
30. Bemis, B. E., Spero, H. J., Bijma, J. & Lea, D. W. Reevaluation of the oxygen isotopic composition of planktonic foraminifera: experimental results and revised paleotemperature equations. *Paleoceanography* 13, 150–160 (1998).

Supplementary Information accompanies the paper on www.nature.com/nature.

Acknowledgements The support of French MENRT, TAAF, CNRS/INSU and IPEV to the Marion-Dufresne and the IMAGES Program was necessary to perform this work. We are grateful to N. Buchet, B. Le Coat, P. Field and S. Perron-Cushman for technical help, and to A. Broccoli, A. C. Ravelo, D. Andreasen and C. Lévi for comments. This study was partially supported by Rutgers' post-doctoral fellowship to T.d.G.-T. and by the National Science Foundation to Y.R.

Competing interests statement The authors declare that they have no competing financial interests.

Correspondence and requests for materials should be addressed to T.d.G.-T. (garidel@marine.rutgers.edu).

Long-term sensitivity of soil carbon turnover to warming

W. Knorr¹, I. C. Prentice^{1,2}, J. I. House^{1,2} & E. A. Holland^{1,3}

¹Max Planck Institute for Biogeochemistry, PO Box 100164, D-07701 Jena, Germany

²QUEST, Department of Earth Sciences, University of Bristol, Wills Memorial Building, Bristol BS8 1RJ, UK

³National Center for Atmospheric Research, PO Box 3000, Boulder, Colorado 80305, USA

The sensitivity of soil carbon to warming is a major uncertainty in projections of carbon dioxide concentration and climate¹. Experimental studies overwhelmingly indicate increased soil organic carbon (SOC) decomposition^{2–8} at higher temperatures, resulting in increased carbon dioxide emissions from soils. However, recent findings have been cited as evidence against increased soil carbon emissions in a warmer world^{9,10}. In soil warming experiments, the initially increased carbon dioxide efflux returns to pre-warming rates within one to three years^{10–14}, and apparent carbon pool turnover times are insensitive to temperature¹⁵. It has already been suggested that the apparent lack of temperature dependence could be an artefact due to neglecting the extreme heterogeneity of soil carbon¹⁶, but no explicit model has yet been presented that can reconcile all the above findings. Here we present a simple three-pool model that partitions SOC into components with different intrinsic turnover rates. Using this model, we show that the results of all the soil-warming experiments are compatible with long-term temperature sensitivity of SOC turnover: they can be explained by

rapid depletion of labile SOC combined with the negligible response of non-labile SOC on experimental timescales. Furthermore, we present evidence that non-labile SOC is more sensitive to temperature than labile SOC, implying that the long-term positive feedback of soil decomposition in a warming world may be even stronger than predicted by global models^{1,17–20}.

The short-term temperature dependence of decomposition and heterotrophic respiration in soils is well established experimentally (see, for example, refs 2–8), and is modelled either by a constant Q_{10} value of ~ 2 (Q_{10} is the proportional increase in reaction rate for a 10 K warming) or more accurately by the Arrhenius equation or slight modifications thereof^{3,5}. This short-term response conceals enormous heterogeneity in SOC: soils include compounds with intrinsic turnover times (that is, turnover times at a reference temperature) ranging from <1 yr to $>6 \times 10^3$ yr (refs 4, 16, 21). Geographical patterns generally show younger SOC ¹⁴C ages in warmer climates^{4,16,21,22}, consistent with the hypothesis that climatic warming should reduce global SOC by reducing residence times.

It is well established that a one-pool representation of SOC dynamics is insufficient to explain experimental findings from warming and incubation experiments^{5,16}. But no model has yet been presented that is able to reconcile the above observations with refs 10–15. To this end, we used experimental data from ref. 6 to fit simple illustrative models of SOC decomposition. The data were derived from temperature-controlled incubations of an undisturbed soil from a tropical rain forest site with a mean annual temperature of 27 °C. Carbon dioxide (CO₂) efflux was measured at ten intervals through a 24-week period (see Methods). The soil was taken to consist of n pools of carbon content c_i , each decaying at a temperature-dependent rate k_i over time t :

$$dc_i(t)/dt = -k_i c_i(t) \quad (1)$$

We considered Arrhenius models with a single reference decay rate A for all pools, and n activation energy values E_i :

$$k_i(T_k) = A \exp(-E_i/RT_k) \quad (2)$$

where T_k is soil temperature in kelvin and A the theoretical decay rate at $E_i = 0$; R is the universal gas constant.

We started by fitting a model with $n = 1$, then increased n until no significant improvement of a χ^2 criterion was obtained (see Methods). This procedure selected a model with $n = 2$ (Table 1a). The time-course of CO₂ efflux during the 24-week observation period is thus appropriately modelled by the decay of two SOC pools following Arrhenius kinetics but differing in activation energy. But the total soil carbon content implied by the two-pool decay models (Table 1) is far less than the measured initial carbon content of 29.4 gC per kg soil. We therefore added a third pool, effectively inert over the timescale of the experiment, which accounts for $\sim 95\%$ of SOC. The fitted turnover time for this pool is arbitrarily large. A lower limit for its activation energy (see Methods) is 68,000 J mol⁻¹, implying a turnover time of 260 yr at 25 °C.

Adopting the three-pool Arrhenius model with $E_3 = 68,000$ J mol⁻¹, we first reconsider the analysis of incubation experiments in ref. 15. A one-pool model based on equation (1) was used there to analyse the fractional decay of SOC after incubation for 1 yr. Inverting the solution to equation (1) for a single SOC pool gives an apparent turnover time, τ :

$$\tau = -t_1 / \ln[c(t_1)/c(0)] \quad (3)$$

where t_1 is the time at the end of the incubation¹⁵. Applying this definition to modelled total soil carbon content $c(t_1)$, with $t_1 = 1$ yr, markedly reduces the apparent sensitivity of turnover time to temperature despite the built-in temperature dependence of all the rate constants (Fig. 1a). This paradox arises because total soil carbon $c(t)$ is dominated by the slowest pool; over a year, the two faster pools have largely decayed. The apparent turnover time is then closely approximated by $c_3(0)/c(0)$. This ratio contains no

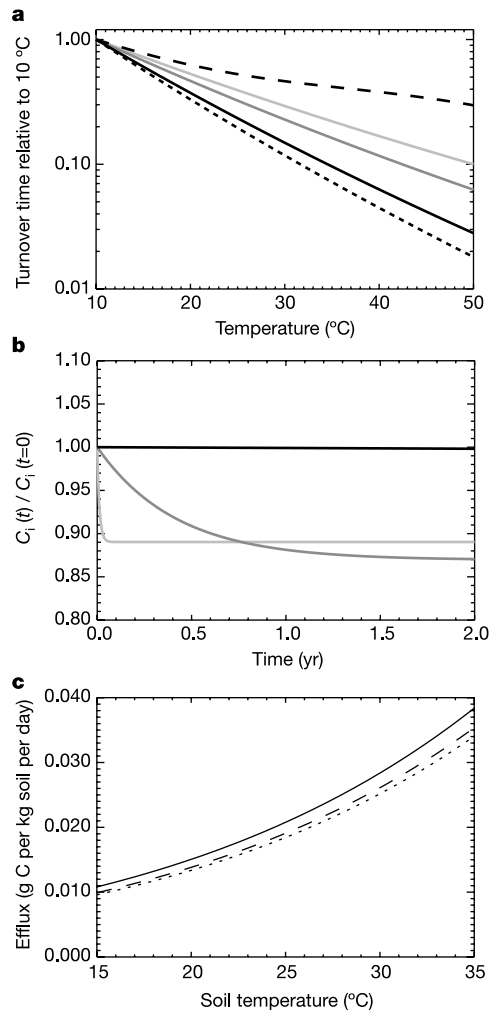


Figure 1 Temperature responses of the three-pool Arrhenius model. **a**, Turnover times of the three pools as a function of temperature (solid lines): fast (light grey), intermediate (medium grey) and slow (black). The dotted black line uses the best-fit value for activation energy (Table 1b), the solid line 68,000 J mol⁻¹ as a lower bound. The uppermost line shows turnover times calculated with the method of ref. 24. **b**, Modelled response of the pool sizes to a 2 °C warming (grey shades as **a**). **c**, Modelled temperature response of CO₂ efflux after warming the soil by 2 °C for 1 month (dashed line) or 1 yr (dotted line), compared to the temperature response without warming (solid line).

information about the temperature sensitivity of the carbon pools.

Responses of the three carbon pools in an idealized soil warming experiment (a step change of +2 °C maintained over two years: see Methods) were then simulated, assuming the soil was initially in equilibrium at 27 °C (Fig. 1b). The first pool is close to a new equilibrium after 1 month and the second pool after 2 yr. CO₂ efflux from the soil is dominated by the first two pools and returns to equilibrium within ~2 yr. The observed equilibration of the faster SOC components has been attributed to a process called ‘acclimation’^{12,13}. Acclimation (or acclimatization) implies a gradual adaptation of living organisms or assemblages to changing environmental conditions: in this case, a change in the temperature sensitivity of soil microbial metabolism. Figure 1c shows the theoretically derived temperature dependence of the SOC decomposition rate before and immediately after warming the soil for different periods. The effect of the warming treatment is to lower the rate at any given temperature. This adjustment is to be expected as a simple result of fast labile SOC depletion relative to the bulk of more stable SOC. No biological adaptation, as suggested by the term ‘acclimation’^{12,13}, is needed to explain it.

Equation (2) implies that the more slowly decaying fractions of SOC should respond more steeply to temperature change than the labile components (Fig. 1a, equation (4), and ref. 23). To test this hypothesis, we used results from ref. 5 where a two-pool Arrhenius model was fitted to data from a wide range of incubation experiments with varying temperature from different climates and soil types. We selected those 13 experiments that lasted at least 100 days (range: 104–720 days). The model of ref. 5 defines only one activation energy, E , for both pools. However, there is a highly significant negative correlation between values of E (in units of J mol⁻¹) estimated in ref. 5 and the initial fraction of the fast pool: $E = 60,578 - [35,556 c_1(0)/c(0)]$, $r = -0.70$, $P = 0.996$. We also computed the total SOC turnover time, defined as $c(0)/[-dc(0)/dt]$, at a reference temperature of 15 °C. A clear trend, with higher E for soils with a higher turnover time, is shown—consistent with predictions of our three-pool model (Fig. 2). Both analyses indicate that soils containing less labile material have a higher sensitivity of decomposition to temperature.

Models of SOC decay in a warming climate are based on experimental temperature responses^{1,17–20}. Experiments lasting months to years measure only the response of the faster decaying components. But most of the Earth’s SOC is in slowly decaying components²¹. Our results suggest that even stronger temperature responses, and a greater magnitude of the positive SOC feedback, are to be expected in response to warming over decades to centuries.

In contrast with earlier studies^{24,25}, ref. 26 showed little change in total SOC along a high-latitude temperature gradient. But the trend in net primary production along this gradient is probably similar to that in SOC turnover, leading to approximately constant SOC (see

Table 1 Results of model optimization

a Arrhenius model					
Model version	Number of parameters	Initial concentration (g C per kg soil)	Turnover time at 25 °C (d)	Activation energy (J mol ⁻¹)	χ^2
One-pool	3	1.07	80.3	44,259	3.49
Two-pool	5	0.08	5.5	43,960	1.18
		1.4	203.8	52,920	
b Arrhenius model with total soil carbon					
Model version	Number of parameters	Initial concentration (g C per kg soil)	Turnover time at 25 °C (d)	Activation energy (J mol ⁻¹)	χ^2
Two-pool	5	0.33	25	48,092	2.95
		29.07	17,232	64,334	
Three-pool	7	0.08	5.5	43,805	1.18
		1.39	201.7	52,744	
		27.93	2.77 × 10 ⁶	76,359	

Optimal parameter and reduced χ^2 values for models fitted to CO₂ efflux data obtained by incubating fractions of the same soil sample at temperatures ranging from 15 to 45 °C. **a**, Arrhenius models; **b**, the same with initial soil carbon content as an additional constraint.

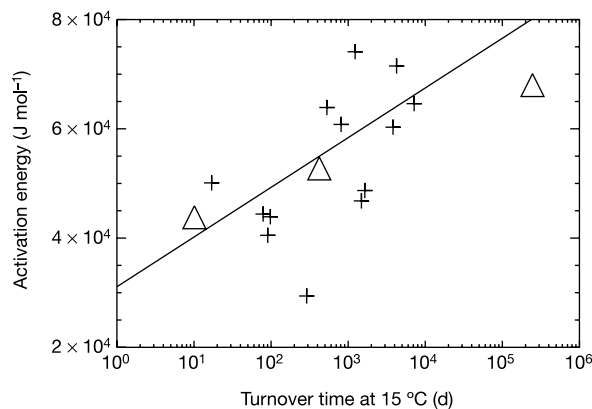


Figure 2 Activation energy and turnover time. Crosses, estimated activation energy for 13 soils from ref. 5 based on incubation experiments and a two-pool Arrhenius model, plotted against inferred turnover time at 15 °C. There is a strong trend for the more slowly cycling pools to have a higher activation energy, implying a steeper temperature response to warming (linear regression: activation energy (J mol^{-1}) = $26,829 + [4,259 \ln(\text{turnover time (d)})]$, $r = 0.59$, $P = 0.98$, solid line). Triangles, comparable values based on the three-pool model of this study.

ref. 27 for a further critique). Total ecosystem respiration shows no pronounced temperature dependence across Europe²⁸ but more closely matches gross primary production, as would be theoretically expected¹: geographical trends in total respiration, unlike ^{14}C ages, convey no information about the temperature dependence of SOC decomposition rates.

That models used to interpret SOC decomposition measurements should account for SOC heterogeneity has been pointed out before^{5,16,29}. Simple models as used here assume that the Arrhenius equation is appropriate to describe any given SOC 'pool', although the pools themselves may be heterogeneous. This simplification is necessary given the limited available experimental data. Using a more complex theoretical model allowing continuous variation in SOC quality, Ågren and Bosatta³⁰ suggested that decomposition rates should vary less when analysed across environments with different 'native' temperatures than when a soil is perturbed away from its native temperature. Our model makes the opposite prediction (see Methods). Both remain to be tested.

Our simple model can account for the apparent lack of temperature dependence presented in ref. 15 as well as the apparent acclimation of total respiration in soil warming experiments^{10–14}. We conclude that the hypothesis of a positive feedback from SOC decay to global warming is fully consistent with the available evidence. Furthermore, the response of soil carbon loss over century timescales is likely to be stronger than is suggested by fixed- Q_{10} models^{17–20}, including coupled climate-carbon-cycle models²⁰. □

Methods

Data

Holland et al.⁶ incubated four replicate homogeneous samples to 10 cm depth of the same soil from an undisturbed forest site in eastern Amazonia (Fazenda Victoria, near Paragominas, State of Pará, Brazil, 2° 59' S, 47° 31' E) at 15 °C, 25 °C, 35 °C, 45 °C and 55 °C for 24 weeks. All samples were kept at field capacity to avoid variations due to soil moisture content. Soil CO_2 efflux was determined through measurement of accumulated CO_2 concentrations in the sampling jars and ventilation before and after the sampling period, which was 1 h during the first week, and 24 h thereafter. Measurements were made during the first day after pre-incubating overnight at the experimental temperature, and then 3 days, 1, 2, 4, 7, 10, 13, 17 and 24 weeks after the pre-incubation. The sample frequency followed the pattern of soil heterotrophic respiration, being more frequent at the beginning when respiration rates were high and less so as respiration declined. During the first day, the linear increase of CO_2 concentrations in the sampling jar was tested by measurements at 1, 3, 5 and 24 h after pre-incubation.

Model fitting

Model parameters were fitted using the downhill simplex method. We computed the reduced χ^2 value to quantify agreement between modelled and measured CO_2 fluxes

assuming a constant measurement error of $0.1 \mu\text{g C per g soil per h}$, based on data given by ref. 6, Fig. 3e. The error bars shown there were computed from the different measured fluxes of four soil samples, and represent both instrumental errors and differences between samples. In order to account only for approximately constant instrument errors, we took a value close to the smallest recorded error bars at the lowest temperature and lowest flux. Incubation data for 55 °C were excluded from the model fitting because they showed evidence of thermal inhibition. The starting time of the model, $t = 0$, was taken to be 12 h before the end of the pre-incubation period, such that the efflux at $t = 1, 3$, or 7 days corresponds to the measured daily average flux during day 1, 3, or 7 of the incubation. Adding more pools than shown in Table 1 produced no significant improvement in χ^2 . When total soil carbon before incubation was used as an additional constraint (Table 1b), a three-pool model was required, but the activation energy of the slowest pool (E_3) was highly uncertain. Therefore, the parameter E_3 was not counted when computing degrees of freedom to derive the reduced χ^2 value. We varied E_3 while holding all other parameters at their optimized values, and obtained reduced χ^2 values of 1.24 at 70,000, 1.36 at 68,000, and 1.94 at 66,000 J mol^{-1} , respectively. $E_3 = 68,000 \text{ J mol}^{-1}$ was adopted as a practical lower bound.

Model application

To mimic soil warming experiments, we assumed that initial concentrations, $c_i(0)$, $i = 1, \dots, n$, were equal to the equilibrium concentrations at the mean annual temperature, T_0 . This assumption implies constant input rates $k_i(T_0)c_i(0)$. When the soil is warmed by an amount ΔT over a time t_1 , with no change in inputs, the sizes of the carbon pools become:

$$c_i(t_1) = c_i(0) \left\{ e^{-k_i(T_0 + \Delta T)t_1} + \frac{k_i(T_0)}{k_i(T_0 + \Delta T)} (1 - e^{-k_i(T_0 + \Delta T)t_1}) \right\} \quad (4)$$

Immediately after the treatment, total CO_2 efflux at any temperature T is computed as $F_{\text{tot}}^{\text{treated}} = \sum k_i(T)c_i(t_1)$, which differs from the initial temperature response $F_{\text{tot}} = \sum k_i(T)c_i(0)$. Because $c_3(0) \gg c_2(0)$ and $c_3(0) \gg c_1(0)$, the response of the soil on long timescales as seen in total SOC pool changes can be approximated by the ratio of initial to warmed decay rates (or warmed to initial turnover times) of the slow pool: $\Sigma c_i(t_1)/\Sigma c_i(0) \approx c_3(t_1)/c_3(0) \approx k_3(T_0)/k_3(T_0 + \Delta T)$. Conversely, because $k_3 \ll k_1$ and $k_3 \ll k_2$, over short timescales total SOC hardly changes, while the flux response is dominated by changes in the fast pools. If we define an average decomposition rate $k = [\Sigma k_i c_i(t_1)]/\Sigma c_i(t_1)$, the inverse of the average turnover time, the variation in k over short timescales is dominated by changes in total efflux, $\Sigma k_i c_i(t_1)$, while changes over long timescales are determined by changes in total SOC, $\Sigma c_i(t_1)$. Since total SOC is dominated by the slow pool, and total efflux by the fast pools, and $E_3 > E_2 > E_1$, we predict a steeper temperature response of k for different soils at varying native temperatures (long-term response), than when the same soils are perturbed away from their native temperatures (short-term response).

Received 26 May; accepted 24 November 2004; doi:10.1038/nature03226.

1. Prentice, I. C. et al. in *Climate Change 2001: The Scientific Basis* (eds. Houghton, J. T. et al.) 185–225 (Cambridge Univ. Press, Cambridge, 2001).
2. Jenkinson, D. S. & Ayanaba, A. Decomposition of C-14-labeled plant material under tropical conditions. *Soil Sci. Soc. Am. J.* **41**, 912–915 (1977).
3. Lloyd, J. & Taylor, J. A. On the temperature dependence of soil respiration. *Funct. Ecol.* **8**, 315–323 (1994).
4. Trumbore, S. E., Chadwick, O. A. & Amundson, R. Rapid exchange between soil carbon and atmospheric carbon dioxide driven by temperature change. *Science* **272**, 393–396 (1996).
5. Kätterer, T., Reichstein, M., Andrén, O. & Lomander, A. Temperature dependence of organic matter decomposition: a critical review using literature data analyzed with different models. *Biol. Fertil. Soils* **27**, 258–262 (1998).
6. Holland, E. A., Neff, J. C., Townsend, A. R. & McKeown, B. Uncertainties in the temperature sensitivity of decomposition in tropical and subtropical ecosystems: Implications for models. *Glob. Biogeochem. Cycles* **14**, 1137–1151 (2000).
7. Dalias, P., Anderson, J. M., Bottner, P. & Coûteaux, M. M. Long-term effects of temperature on carbon mineralisation processes. *Soil Biol. Biochem.* **33**, 1049–1057 (2001).
8. Sanderman, J., Amundson, R. & Baldocchi, D. D. Application of eddy covariance measurements to the temperature dependence of soil organic matter mean residence time. *Glob. Biogeochem. Cycles* **17**, 1061–1075 (2003).
9. Grace, J. & Rayment, M. Respiration in the balance. *Nature* **404**, 819–820 (2000).
10. Jarvis, P. & Linder, S. Constraints to growth of boreal forests. *Nature* **405**, 904–905 (2000).
11. Peterjohn, W. T. et al. Responses of trace gas fluxes and N availability to experimentally elevated soil temperatures. *Ecol. Appl.* **4**, 617–625 (1994).
12. Oechel, W. C. et al. Acclimation of ecosystem CO_2 exchange in the Alaskan Arctic in response to decadal climate warming. *Nature* **406**, 978–981 (2000).
13. Luo, Y. Q., Wan, S. Q., Hui, D. F. & Wallace, L. L. Acclimatization of soil respiration to warming in a tall grass prairie. *Nature* **413**, 622–625 (2001).
14. Rustad, L. E. et al. A meta-analysis of the response of soil respiration, net nitrogen mineralization, and aboveground plant growth to experimental ecosystem warming. *Oecologia* **126**, 543–562 (2001).
15. Giardina, C. P. & Ryan, M. G. Evidence that decomposition rates of organic carbon in mineral soil do not vary with temperature. *Nature* **404**, 858–861 (2000).
16. Davidson, E. A., Trumbore, S. E. & Amundson, R. Biogeochemistry—Soil warming and organic carbon content. *Nature* **408**, 789–790 (2000).
17. Jenkinson, D. S., Adams, D. E. & Wild, A. Model estimates of CO_2 emissions from soil in response to global warming. *Nature* **351**, 304–306 (1991).
18. Cao, M. K. & Woodward, F. I. Dynamic responses of terrestrial ecosystem carbon cycling to global climate change. *Nature* **393**, 249–252 (1998).
19. Cramer, W. et al. Global response of terrestrial ecosystem structure and function to CO_2 and climate change: results from six dynamic global vegetation models. *Glob. Change Biol.* **7**, 357–374 (2001).
20. Cox, P. M., Betts, R. A., Jones, C. D., Spall, S. A. & Totterdell, I. J. Acceleration of global warming due to carbon-cycle feedbacks in a coupled climate model. *Nature* **408**, 184–187 (2000).

21. Trumbore, S. Age of soil organic matter and soil respiration: Radiocarbon constraints on belowground C dynamics. *Ecol. Appl.* **10**, 399–411 (2000).
22. Bird, M. I., Chivas, A. R. & Head, J. A latitudinal gradient in carbon turnover times in forest soils. *Nature* **381**, 143–146 (1996).
23. Bosatta, E. & Ågren, G. I. Soil organic matter quality interpreted thermodynamically. *Soil Biol. Biochem.* **31**, 1889–1891 (1999).
24. Jenny, H. Relation of temperature to the amount of nitrogen in soils. *Soil Sci.* **27**, 169–188 (1929).
25. Burke, I. C. *et al.* Texture, climate, and cultivation effects on soil organic matter content in the US grassland soils. *Soil Sci. Soc. Am. J.* **53**, 800–805 (1989).
26. Liski, J., Ilvesniemi, H., Makela, A. & Westman, C. J. CO₂ emissions from soil in response to climatic warming are overestimated—The decomposition of old soil organic matter is tolerant of temperature. *Ambio* **28**, 171–174 (1999).
27. Ågren, G. I. Temperature dependence of old soil organic matter. *Ambio* **29**, 55 (2000).
28. Janssens, I. A. *et al.* Productivity overshadows temperature in determining soil and ecosystem respiration across European forests. *Glob. Change Biol.* **7**, 269–278 (2001).
29. Gu, L., Post, W. M. & King, A. W. Fast labile carbon turnover obscures sensitivity of heterotrophic respiration from soil to temperature: A model analysis. *Glob. Biogeochem. Cycles* **18**, 1022–1032 (2004).
30. Ågren, G. I. & Bosatta, E. Reconciling differences in predictions of temperature response of soil organic matter. *Soil Biol. Biochem.* **34**, 129–132 (2002).

Acknowledgements We thank A. Arneth, F. Badeck, T. Christensen, G. Churkina, C. Czimczik, C. Gracia, V. Hahn, E. Hobbie, J. Kaplan, F. Joos, J. Liski, J. Lloyd, S. Sabaté, D. Schimel, U. Seibt, S. Sitch, P. Smith, J. Trenbath, S. Trumbore and R. Valentini for discussions, and S. Schott for technical editing.

Competing interests statement The authors declare that they have no competing financial interests.

Correspondence and requests for materials should be addressed to W.K. (wknorr@bgc-jena.mpg.de).

Early Pliocene hominids from Gona, Ethiopia

Sileshi Semaw¹, Scott W. Simpson^{2,3}, Jay Quade⁴, Paul R. Renne^{5,6}, Robert F. Butler⁴, William C. McIntosh^{7,8}, Naomi Levin⁴, Manuel Dominguez-Rodrigo⁹ & Michael J. Rogers¹⁰

¹CRAFT Stone Age Institute, Indiana University, 1392 West Dittmore Road, Gosport, Indiana, 47433-9531, USA

²Department of Anatomy, Case Western Reserve University-School of Medicine, 10900 Euclid Avenue and ³Laboratory of Physical Anthropology, Cleveland Museum of Natural History, Cleveland, Ohio, 44106, USA

⁴Department of Geosciences, University of Arizona, Tucson, Arizona, 85721, USA

⁵Berkeley Geochronology Center, 2455 Ridge Road and

⁶Department of Earth and Planetary Science, University of California, Berkeley, California, 94709, USA

⁷New Mexico Bureau of Geology and Mineral Resources and

⁸Department of Earth and Environmental Science, New Mexico Institute of Mining and Technology, Socorro, New Mexico, 87801, USA

⁹Departamento de Prehistoria y Arqueología, Facultad de Geografía e Historia, Universidad Complutense de Madrid, Ciudad Universitaria s/n, 28040 Madrid, Spain

¹⁰Department of Anthropology, Southern Connecticut State University, 501 Crescent Street, New Haven, Connecticut, 06515-1355, USA

Comparative biomolecular studies suggest that the last common ancestor of humans and chimpanzees, our closest living relatives, lived during the Late Miocene–Early Pliocene^{1,2}. Fossil evidence of Late Miocene–Early Pliocene hominid evolution is rare and limited to a few sites in Ethiopia^{3,4,5}, Kenya⁶ and Chad⁷. Here we report new Early Pliocene hominid discoveries and their palaeoenvironmental context from the fossiliferous deposits of As Duma, Gona Western Margin (GWM), Afar, Ethiopia. The hominid dental anatomy (occlusal enamel thickness, absolute and relative size of the first and second lower molar crowns, and premolar crown and radicular anatomy) indicates attribution to *Ardipithecus ramidus*. The combined radioisotopic and palaeomagnetic data suggest an age of between 4.51 and 4.32 million

years for the hominid finds at As Duma. Diverse sources of data (sedimentology, faunal composition, ecomorphological variables and stable carbon isotopic evidence from the palaeosols and fossil tooth enamel) indicate that the Early Pliocene As Duma sediments sample a moderate rainfall woodland and woodland/grassland.

The Early Pliocene fossiliferous sediments at As Duma in the Gona Palaeoanthropological Research Project (GPRP) area are located in the Busidima and Gawis drainages at the base of the Gona western escarpment in the Afar portion of the Ethiopian rift (Fig. 1). During four field seasons between 1999 and 2003, over 1,500 fossils have been recovered from 40 palaeontological sites, seven of which included hominid remains (Table 1 and Supplementary Table 1). These fossil-rich deposits are divided from the younger Hadar and Busidima Formations (3.4 million years (Myr) ago to <0.5 Myr ago) to the east by a major normal fault (Fig. 1). Most of the As Duma sites are distributed across two distinct fault blocks separated by an apparently minor normal fault. The west-dipping orientation of the minor fault implies that all of the GWM-3 block (containing sites GWM-3/3w, and -16) is stratigraphically lower than the GWM-5 block (sites GWM-1, -5m, -5sw and -9), although the lack of stratigraphic repetition between blocks makes the time separation between sites and blocks unclear. Given their many sedimentologic similarities, we interpret these blocks as a part of a single depositional sequence. Similarities

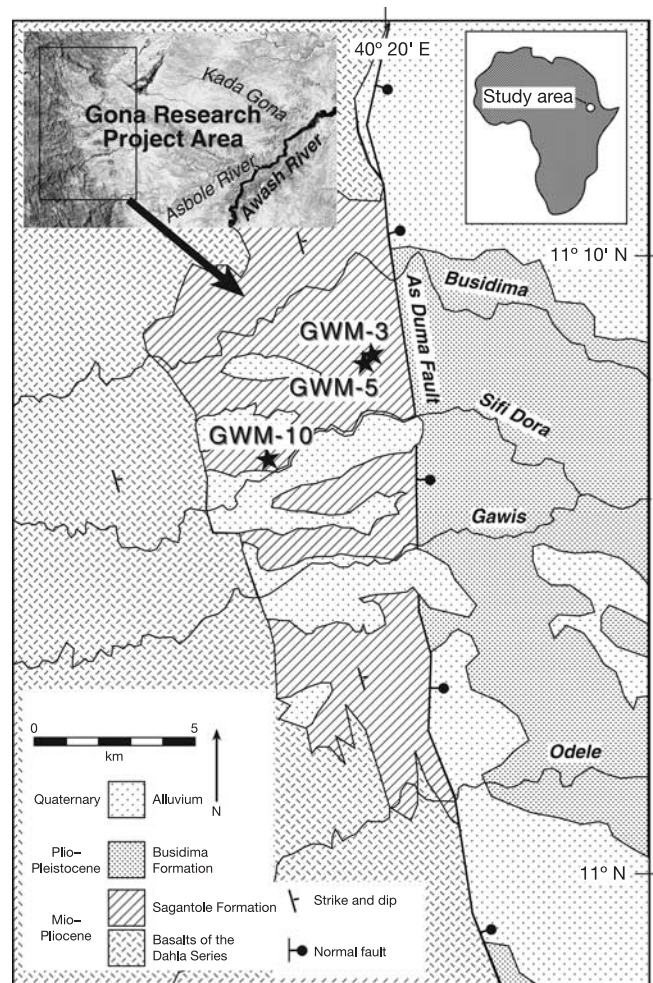


Figure 1 A geological map and thematic mapper (TM) image insert depicting the location of the hominid sites at As Duma in the Gona Western Margin, GPRP study area, Afar, Ethiopia. The north–south trending As Duma fault offsets the younger (≤ 3.4 Myr ago) Hadar and Busidima Formations against the newly identified Early Pliocene deposits of the Sagantole Formation.

that distinguish the GWM-3 and GWM-5 deposits from the underlying Dahla Series basalts and overlying Hadar Formation include: the same mix of small-scale fluvial, lacustrine and locally volcanoclastic deposition; similar colour and induration; and the similar petrographic character of the fluvial sands. The GWM-3 site stratigraphic section (Fig. 2a) is capped by a laterally extensive tuff of spring origin containing gastropods, fossilized wood fragments, *Celtis* cf. *africana* seeds and vertebrate fossils, including hominids. At the base of the GWM-5 stratigraphic block are dark-grey mudstones resting directly on a minor basalt flow (Fig. 2b). The fossiliferous mudstone, which has yielded hominids, contains abundant tuffaceous and volcanoclastic material interpreted to represent ongoing volcanic eruptions into a shallow marsh that lapped onto the basalt flow.

Dates from two different crystal-rich tuffs (GONASH-51 and GONASH-52) found in the middle of the lower lacustrine interval ~14 m below GWM-3 and from the stratigraphically higher basalt flow under GWM-5sw indicate that the deposits are Early Pliocene in age (Fig. 2a, b). The two tuffs were dated by laser fusion $^{40}\text{Ar}/^{39}\text{Ar}$ analysis of single plagioclase crystals using methods described previously⁸ (Supplementary Figs 1 and 2). Excluding the obvious xenocrysts, 23 of 23 (GONASH-51) and 16 of 18 (GONASH-52) plagioclase crystals yielded statistically homogeneous populations with weighted mean ages ($\pm 2\sigma$) of 4.56 ± 0.23 Myr and 4.59 ± 0.45 Myr, respectively. Although the individual crystal ages

are relatively scattered (unweighted population means and standard deviations of 4.45 ± 0.53 Myr ($n=23$) and 4.31 ± 0.91 Myr ($n=15$) for G-51 and G-52, respectively) their intra-sample homogeneity and variable precision justify the weighted mean as the best estimate for the ages (Supplementary Figs 1 and 2; Supplementary Table 2).

Two samples of groundmass concentrate (WMASH-25 and -27) from the basalt flow beneath GWM-5 were dated by resistance-furnace incremental heating yielding plateau ages of 4.17 ± 0.21 Myr and 4.06 ± 0.39 Myr, with a weighted mean of 4.15 ± 0.18 Myr (Supplementary Figs 3 and 4; Supplementary Table 3). Palaeomagnetic data (referred to the timescale in ref. 9, recalculated for conformity with the standard age used in $^{40}\text{Ar}/^{39}\text{Ar}$ dating herein) provide additional constraints on the age of the deposits. The reversed polarity sediments and flows at all sites within these densely sampled stratigraphic sections may have been deposited during a single interval of reversed geomagnetic polarity, given the strong sedimentologic similarities between the GWM-3 and GWM-5 deposits. If this proves correct, the polarity interval that simultaneously satisfies all palaeomagnetic and radioisotopic data is chron C3n.1r with age limits 4.51–4.32 Myr (Fig. 2c). An alternative—but in our view geologically less plausible—interpretation is that GWM-5 is much younger than GWM-3. In this case, the magnetically reversed sediments of GWM-3 belong to chron C3.2r (4.83–4.65 Myr) or chron C3.1r (4.51–4.32 Myr), whereas the magnetically reversed sediments of GWM-5 could conceivably belong to chron C2Ar (4.21–3.61 Myr). In this model, palaeomagnetic constraints would place the GWM-3 block between 4.83 and 4.32 Myr ago, and combined radioisotopic results would place the GWM-5 block between 4.21 and 3.61 Myr ago. Further sampling may clarify the chronological position of GWM-5. The Early Pliocene age for all of these localities is supported by the presence of biochronologically useful species recovered from the hominid sites (for example, *Nyanzachoerus jaegeri*¹⁰, *Kolpochoerus deheinzlini*¹¹, *Anancus* cf. *kenyensis*, *Pliopapio alemui*¹² and *Kuseracolobus aramisi*¹²). Another fossil hominid site, GWM-10, although not yet dated geochronologically, is biochronologically indistinguishable from other As Duma localities.

Fossils from at least nine hominid individuals were recovered from the As Duma deposits (Fig. 3 and Table 1). No hominid fossil shows evidence of transport. Primarily jaws and teeth fossils, the collection includes three variably complete manual phalanges and a

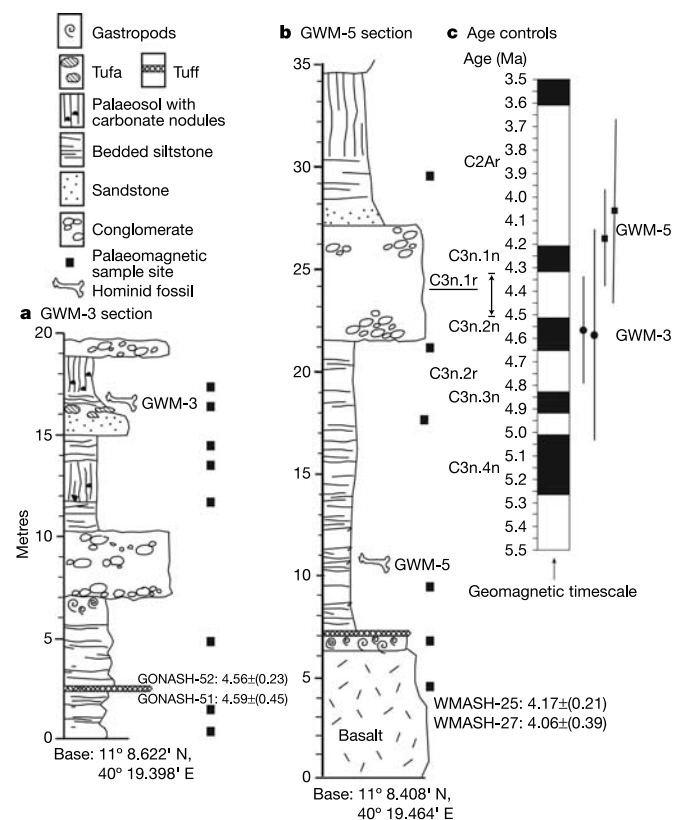


Figure 2 Stratigraphy of the GWM-3 (a) and GWM-5 (b) sites. All of the GWM-3 block in a lies stratigraphically below the minor basalt at the base of the GWM-5 block in b. GPS coordinates based on WGS84 datum. c, Palaeomagnetic timescale (solid, normal polarity; open, reversed polarity) in millions of years but recalculated for conformity with the standard age used in $^{40}\text{Ar}/^{39}\text{Ar}$ dating herein) and $^{40}\text{Ar}/^{39}\text{Ar}$ ages with 2σ uncertainties (solid circles, tuffs from the GWM-3 fault block; solid squares, from basalt at the base of the GWM-5 block). All sediments in both fault blocks are reversely magnetized. C3n.1r (4.32–4.51 Myr) is the only reversed interval where both $^{40}\text{Ar}/^{39}\text{Ar}$ ages also overlap at 2σ uncertainty.

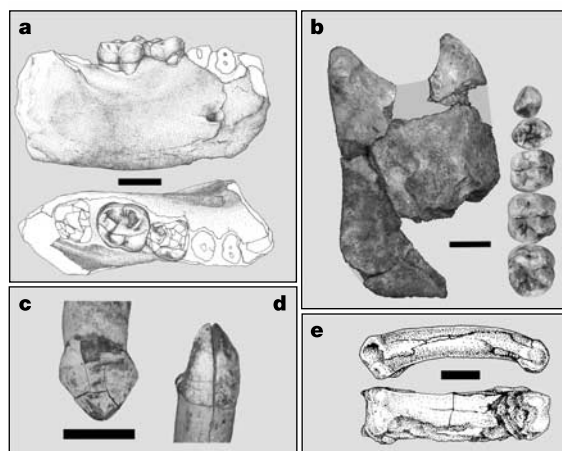


Figure 3 Early Pliocene hominid fossils from As Duma, Gona Western Margin. a, GWM3/P1, lateral and occlusal view of right mandibular corpus. b, GWM5sw/P56, mesial view of left mandibular ramus and occlusal views of right P₃–M₃. Grey shaded area of ramus is area of reconstruction. c, GWM9n/P51, labial view of maxillary left canine. d, GWM9n/P50, labial view of mandibular right canine. e, GWM10/P1, lateral and palmar view of manual proximal phalanx. Solid bar, 1 cm. Drawings by L. Gudž.

Table 1 Early Pliocene hominids from As Duma, Gona, Afar, Ethiopia

Specimen	Element	Finder	Date	GPS-N*	GPS-E*	Measurements (mm)†
GWM1/P37	Proximal pedal phalanx fragment	S. W. Simpson	16 Feb 2001	11° 08.386'	40° 19.597'	Base: DV = 7.8; ML = 9.2
GWM3/P1	Right mandible corpus fragment	Abdu Mohamed Ali	6 Mar 1999	11° 08.613'	40° 19.492'	RM ₁ : MD = 10.2; BL _{MAX} = (10.0); RM ₂ : MD = 12.7; BL _{MAX} = 11.6
GWM3/P2	Proximal manual phalanx fragment	Excavation	7 Mar 1999	11° 08.613'	40° 19.492'	Base: DV = 11.6; ML = 15.2
GWM3W/P185	Mandibular left M ₁	Asahmad Humet	24 Feb 2003	11° 08.640'	40° 19.449'	LM ₁ : MD = (10.5); BL _{MAX} = 10.25
GWM5M/P55	Maxillary left P ⁴	Asahmad Humet	28 Feb 2001	11° 08.657'	40° 19.238'	LP ⁴ : MD = 7.25; BL = 10.65
GWM5SW/P56	Partial dentition	Asahmad Humet	17 Feb 2001	11° 08.547'	40° 19.238'	LI ₂ : MD = 6.65; LL = -; RI ₂ : MD = 6.60; LL = -; L _C & R _C : -; RP ₃ : MD = (8.0), BL = -; LP ₃ : MD = 7.8, BL = 8.8; RP ₄ : MD = 8.0, BL = 9.3; LP ₄ : MD = 7.6, BL = 9.3; RM ₁ : MD = 10.6, BL _{MAX} = 10.7; LM ₁ : MD = 10.8; BL _{MAX} = 11.0; RM ₂ : MD = 12.7; BL _{MAX} = 11.7; LM ₂ : MD = 12.7; BL _{MAX} = 11.7; RM ₃ : MD = 12.8, BL _{MAX} = 11.3; LM ₃ : MD = -, BL _{MAX} = 11.4; U ^C : -; LP ³ : MD = -, BL = 9.2; LM ¹ : MD = 9.7, BL = -
GWM5SW/P58	Intermediate manual phalanx fragment	Excavation	21 Feb 2001	11° 08.547'	40° 19.238'	
GWM9N/P50	Mandibular dentition	Asahmad Humet	11 Feb 2003	11° 08.449'	40° 19.427'	R _C : MD = 8.5, BL = 9.7; RP ₄ : MD = (7.3), BL = 8.7; RM ₁ : MD = 10.7, BL _{MAX} = 10.0; LM ₁ : MD = 10.6; BL _{MAX} = (9.7); LM ₂ : MD = (12.4); BL _{MAX} = 11.3
GWM9N/P51	Maxillary dentition	Kampiro Qairento	11 Feb 2003	11° 08.449'	40° 19.427'	RI ² : MD = 7.2; LL = 6.8; L ^C : MD = 11.1; LL = 10.6; RP ³ : MD = 7.5, BL = 11.0; RP ⁴ : MD = 7.4, BL = 11.0; RM ¹ : MD = 10.5, BL _{MAX} = 12.3; RM ³ : MD = -, BL _{MAX} = -; LM ³ : MD = 11.4, BL _{MAX} = 14.2
GWM10/P1	Proximal manual phalanx	Abdu Mohamed Ali	22 Feb 2000	11° 07.489'	40° 18.112'	Length = 49.9; Base: DV = 11.7; ML = (13.4)
GWM16/P10	Mandibular right M ₃	Asahmad Humet	18 Feb 2003	11° 08.783'	40° 19.058'	RM ₃ : MD = 12.55, BL _{MAX} = (11.9)

*GPS coordinates based on WGS84 datum.

†All measurements are in millimetres. Parentheses indicate that values in measurements are estimates. MD, mesio-distal length; BL_{MAX}, maximum bucco-lingual breadth. DV, dorso-ventral; ML, medio-lateral; LL, labio-lingual; -, no data; subscripts indicate lower teeth, superscripts upper teeth; R, right; L, left; C, canine; I, incisor; M, molar; P, premolar.

pedal phalanx fragment. In labial view, the maxillary canine crown is diamond shaped, with the mesial and distal shoulders about equidistant from the cervix and cusp. The third upper molar (M³) is relatively transversely broad with low relief and complex occlusal anatomy. The lower canine is tall with an asymmetric labial profile. The mesial shoulder is about mid-height on the crown with a steep apical mesial margin (Fig. 3d). A distal marginal tubercle is present and prominent. The major axis of the narrow oval mandibular canine root is more anteriorly oriented rather than more obliquely directed to the tooth row as in later hominids. The unicuspid third lower premolars (P₃s) have an unbifurcated root with two pulp canals (Tome's root). The P₄ crown is mesiodistally compressed with an oval plan, a simple cusp pattern and a single root matching the Aramis (Ethiopia) specimens of *Ardipithecus ramidus*⁵ and distinct from the triangular/quadrangular cervical cross-section of *Australopithecus* multi-rooted P₄s.

The small lower first molars are bunodont with a simple cuspal pattern. The molar buccal margins vary in their degree of occlusal slope from near vertical to oblique. The naturally fractured GWM3/P1 M₁ has entoconid occlusal enamel similar in thickness (1.1 mm) to Aramis *Ar. ramidus*. The rectangular M₃s are distally rounded with complex occlusal topography (additional cusps and accessory occlusal folding) and only moderate cuspal relief. The wear gradient across the molar row is minor, which, along with a molar microwear pattern of numerous long bucco-lingual (BL) oriented striae with few pits, indicates a low abrasion diet.

The M₃ crown is separated from the low ramus by a narrow paramolar sulcus. The mandibular corpora are low and broad. The antero-superiorly opening mental foramen is located midcorpus between the premolars. The M₃-C tooth row in GWM-3/P1 is

sigmoidal with the M₃ lateral to the M₁-M₂ axis and the canine medial to it. Reconstruction of this small, perhaps female, mandible indicates that it is 'V'-shaped, common in smaller hominid mandibles (for example, AL 288-1) and different from the parallel tooth rows characteristic of *Australopithecus anamensis*¹³. The large, diamond shaped maxillary canines, small first molar crowns, thickness of occlusal enamel and morphology of the premolar crown and roots indicate assignment to *Ar. ramidus*.

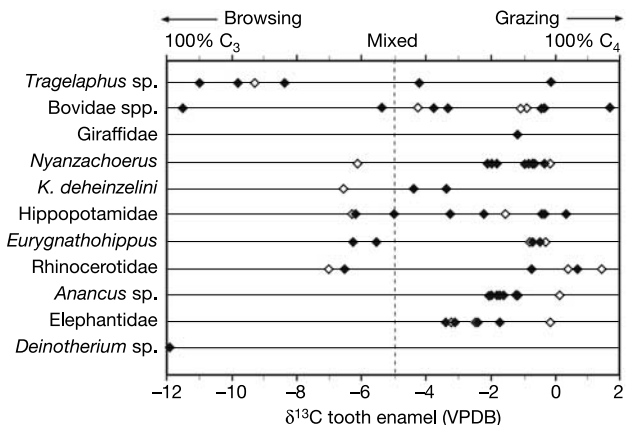


Figure 4 $\delta^{13}\text{C}$ values (GWM-5 block fauna, closed symbols; GWM-3 block fauna, open symbols) of fossil tooth enamel grouped by taxon. *Tragelaphus* sp. = *Tragelaphus* cf. *kyaloae*; Bovidae spp. include multiple species (*Ugandax* sp., Bovini, Reduncini and Bovidae gen. et sp. indet.); *Nyanzachoerus* *jaegeri* = *Nyanzachoerus* cf. *jaegeri* and *Nyanzachoerus* sp.; *K. deheinzellini* = *Kolpochoerus deheinzellini*.

GWM-10/P1 is a nearly complete manual left proximal phalanx probably from ray IV. The GWM-3/P2 manual proximal phalanx, represented by the proximal ~40% of the bone, is large. These phalanges are similar in overall morphology to those from Hadar, with two differences: a more transversely broad proximal articular surface and longer absolute length¹⁴. The pedal proximal phalanx fragment retains the proximal third of the bone. The transversely broad oval proximal facet is oriented dorsally, a character diagnostic of bipedality and a trait also seen in the latest Miocene hominid *Ardipithecus kadabba*^{3,4}.

Faunal composition, bovid tooth mesowear and stable carbon isotopic composition of tooth enamel and soil carbonate provide palaeoenvironmental indicators for As Duma. *N. jaegeri*, a grazer considered to be ancestral to the more hypsodont *Notochoerus* suids^{10,15}, is the most common suid whereas *Nyanzachoerus kanamensis* and the mixed feeding *K. deheinzlini* are either rare or absent at individual hominid sites. The papionine monkey *Pliopapio alemui* is more common (~65%) than the colobine *Kuseracolobus aramisi* (~35%). Observations of the bovid maxillary molar mesowear¹⁶ shows that the tragelaphines (~38% of all bovid individuals) have the characteristic wear pattern of browse or mixed (browse and graze) feeders. The predominance of low, rounded cusps in the remaining bovids indicates a grazing diet, agreeing with the isotopic composition of their teeth (see below). Fossils of the mole rat (*Tachyoryctes* sp.), currently found in montane grasslands, perhaps provide evidence of higher elevation in the past.

Most palaeosols at As Duma are calcareous vertisols that form in seasonal, low-rainfall (<~1,000 mm yr⁻¹) tropical environments^{17,18}. Fifteen carbonate nodules were sampled from ten palaeosols at and around (2 km radius) GWM-3. Sixty-eight total analyses yielded a range of $\delta^{13}\text{C}$ values (VPDB; Vienna Pee Dee belemnite standard) from -3.9‰ to -12.0‰, for an average of -7.5‰ indicating an environment of woodlands and grassy woodlands reflecting a mixture dominated by C_3 (trees and shrubs) and fewer C_4 grasses (warm-season grasses), assuming pure end-member carbonate $\delta^{13}\text{C}$ values of -12‰ and +2‰, respectively¹⁹ (Supplementary Table 4). Only locally, particularly in the vicinity of GWM-3, do analyses indicate the presence of up to 50% grasses, where values range between -3.9‰ to -7.9‰ and average -6.0‰ ($n = 4$).

Stable carbon isotope values in enamel were obtained for most of the macrofauna (except primates) found at the hominid sites (Fig. 4 and Supplementary Table 5). $\delta^{13}\text{C}$ values fall between -11.9‰ and +1.7‰ ($n=73$), reflecting a range of diet composed of C_3 ($\delta^{13}\text{C} = -12$ ‰) and C_4 plants ($\delta^{13}\text{C} = +2$ ‰). The diets of most of these large mammals indicate a significant reliance on C_4 grasses ($\delta^{13}\text{C}_{\text{enamel}} > -5$ ‰, $n=57$). Taxa with diets dominated by C_4 plants (grazers) include *N. jaegeri*, *Hexaprotodon harvardi*, most rhinocerotidae, *Anancus* sp., most bovids, all Elephantidae, most *Eurygnathohippus* sp., and a sivathere, which all returned $\delta^{13}\text{C}$ values > -5.0 ‰. *K. deheinzlini*, some rhinocerotidae and some *Eurygnathohippus* produced intermediate values. A *Deinotherium* and most tragelaphines yielded $\delta^{13}\text{C}$ values typical of a primarily C_3 (browsing) diet.

The As Duma Early Pliocene deposits sample a low-relief, internally draining basin with lakes, swamps, springs and streams amid local volcanic centres. The carbonate-rich palaeosols formed in a sub-humid and seasonally dry climate in a landscape covered by C_3 -dominated woodland and grassy woodland. At As Duma, *Ardipithecus* was exposed to this mosaic of environments. Only further evidence that more tightly associates the hominids (and other fauna) with the varied local environments will reveal the habitat preferences of these early hominids. □

Methods

⁴⁰Ar/³⁹Ar dating

⁴⁰Ar/³⁹Ar dating of single plagioclase crystals separated from GONASH-51 and -52 used the facilities at the Berkeley Geochronology Center (California, USA) following

procedures described by ref. 20 (see Supplementary Information). Plagioclases from GONASH-51 and -52 have high Ca/K (24–65 and 6–13, respectively) that consequently yield relatively imprecise single crystal ages. The plagioclase single crystal ages are intrinsically less precise than the basalt ages due to their low K and high Ca concentrations coupled with the imperative need (demonstrated by refs 8, 21) to analyse crystals individually in order to recognize xenocrystic contamination. The low precision of individual plagioclase analyses unfortunately only enables recognition of xenocrysts that are substantially older, whereas xenocrysts as much as 0.5 Myr older might go undetected in the present data. Furthermore, these low-K plagioclases are relatively susceptible to subtle excess argon contamination. These possible sources of geologic error, not reflected in the analytical precision, both would bias the results towards spuriously old ages. The possibility of xenocrystic contamination precludes analysis of multiple crystal aliquots; indeed, two plagioclase crystals (with anomalously low Ca/K) from GONASH-52 yielded ages of 334 ± 2 Myr and 8.00 ± 1.06 Myr (1 σ errors), illustrating the necessity of single-crystal analysis.

Groundmass concentrates from basalt samples WMASH-25 and -27 were analysed at New Mexico Geochronology Research Lab (Socorro, New Mexico) following full procedures described in ref. 22. Reported ages are based on 28.02 Myr for the Fish Canyon sanidine standard²⁰, which yields ages approximately 0.7% older than the calibration used in refs 8 and 23, and errors are reported at the 2 σ level. Electron microprobe examination of the two basalt samples showed some clay alteration and dissolution of groundmass and phenocryst phases. However, remaining K_2O was concentrated in unaltered plagioclase, and the relatively flat age spectra and moderately high radiogenic yields suggest that alteration phases degassed during low temperature steps, and that the weighted mean ages of higher temperature steps are an accurate record of eruption age of the basalts.

Palaeomagnetism

Four oriented samples for palaeomagnetic analysis were collected from each of 14 sites in the GWM-3 block and eight sites in the GWM-5 block. All samples were thermally demagnetized in ≥ 12 temperature steps ranging up to 580 °C, and characteristic remnant magnetization (ChRM) directions for each sample were determined by principal component analysis²⁴. Site mean ChRM directions were calculated using Fisher statistics²⁵. All site-mean ChRM directions from continuous sections in both blocks are reverse polarity.

Isotope analysis (see Supplementary Information)

Soil carbonate nodules were roasted under vacuum at 400–430 °C before conversion to CO_2 with 100% phosphoric acid. Micro-sampling of individual nodules avoided mm-scale secondary calcite veins that cross-cut many nodules. Multiple $\delta^{13}\text{C}$ analyses of single nodules in individual palaeosols tend to cluster ± 1.0 ‰. Crushed enamel was reacted with 1 M CH_3COOH and with 2% H_2O_2 for 1 h. Isotopic analyses were performed on a Finnigan Delta-S and a Finnigan MAT 252 gas-source mass spectrometer at the University of Arizona. Results are presented in the usual δ notation as the per mil (‰) deviation of the sample CO_2 from the PDB standard, where $R = ^{13}\text{C}/^{12}\text{C}$ or $^{18}\text{C}/^{16}\text{C}$, and $\delta = (R_{\text{sample}}/R_{\text{standard}} - 1) (\times 1,000)$.

Received 16 January; accepted 8 November 2004; doi:10.1038/nature03177.

- Chen, F. C. & Li, W. H. Genomic divergences between humans and other hominoids and the effective population size of the common ancestor of humans and chimpanzees. *Am. J. Hum. Genet.* **68**, 444–456 (2001).
- Salem, A.-H. et al. Alu elements and hominid phylogenetics. *Proc. Natl Acad. Sci. USA* **100**, 12787–12791 (2003).
- Haile-Selassie, Y. Late Miocene hominids from the Middle Awash, Ethiopia. *Nature* **412**, 178–181 (2001).
- Haile-Selassie, Y., Suwa, G. & White, T. D. Late Miocene teeth from Middle Awash, Ethiopia, and early hominid dental evolution. *Science* **303**, 1503–1505 (2004).
- White, T. D., Suwa, G. & Asfaw, B. *Australopithecus ramidus*, a new species of early hominid from Aramis, Ethiopia. *Nature* **371**, 306–312 (1994).
- Senut, B. et al. First hominid from the Miocene (Lukeino Formation, Kenya). *C. R. Acad. Sci. Paris* **332**, 137–144 (2001).
- Brunet, M. et al. A new hominid from the Upper Miocene of Chad, Central Africa. *Nature* **418**, 145–151 (2002).
- Renne, P. R., WoldeGabriel, G., Hart, W. K., Heiken, G. & White, T. D. Chronostratigraphy of the Miocene-Pliocene Sagantole Formation, Middle Awash Valley, Afar Rift, Ethiopia. *Bull. Geol. Soc. Am.* **111**, 869–885 (1999).
- Cande, S. C. & Kent, D. V. Revised calibration of the geomagnetic polarity timescale for the Late Cretaceous and Cenozoic. *J. Geophys. Res.* **100**, 6093–6095 (1995).
- White, T. D. in *Paleoclimate and Evolution, with Emphasis on Human Origins* (ed. Vrba, E.) 369–384 (Yale Univ. Press, New Haven, 1995).
- Brunet, M. & White, T. D. Deux nouvelles espèces de Suini (Mammalia, Suidae) du continent africain (Éthiopie, Tchad). *C. R. Acad. Sci. Paris* **332**, 51–57 (2001).
- Frost, S. R. New Early Pliocene Cercopithecidae from Aramis, Middle Awash Valley, Ethiopia. *Am. Mus. Novit.* **3350**, 1–36 (2001).
- Ward, C. V., Leakey, M. G. & Walker, A. C. Morphology of *Australopithecus anamensis* from Kanapoi and Allia Bay, Kenya. *J. Hum. Evol.* **41**, 255–368 (2001).
- Bush, M. E., Lovejoy, C. O., Johanson, D. C. & Coppens, Y. Hominid carpal, metacarpal, and phalangeal bones recovered from the Hadar Formation: 1974–1977 collections. *Am. J. Phys. Anthropol.* **57**, 651–677 (1982).
- Harris, J. M. & Leakey, M. G. in *Lothagam* (eds Leakey, M. G. & Harris, J. M.) 485–519 (Columbia Univ. Press, New York, 2003).
- Fortelius, M. & Solounias, N. Functional characterization of ungulate molars using the abrasion-attrition wear gradient: A new method for reconstructing palaeodiets. *Am. Mus. Novit.* **3301**, 1–36 (2000).
- Dudal, R. Dark gray soils of tropical and sub-tropical regions. *FAO Agric. Dev. Pap.* **83**, 161 (1965).

18. Ahmad, N. in *Pedogenesis and Soil Taxonomy II. The Soil Orders. Developments in Soil Science 11B* (eds Wilding, L. P., Smeck, N. E. & Hall, G. F.) 91–123 (Elsevier, New York, 1983).
19. Cerling, T. E. & Quade, J. in *Continental Indicators of Climate* (eds Swart, P., McKenzie, J. A. & Lohman, K. C.) 217–231 (AGU monogr. 78, American Geophysical Union, Washington DC, 1993).
20. Renne, P. R. *et al.* Intercalibration of standards, absolute ages and uncertainties in $^{40}\text{Ar}/^{39}\text{Ar}$ dating. *Chem. Geol. (Isot. Geosci. Sect.)* 1–2, 117–152 (1998).
21. Semaw, S. *et al.* 2.5 million-year-old stone tools from Gona, Ethiopia. *Nature* 385, 333–338 (1997).
22. McIntosh, W. C. & Chamberlin, R. M. $^{40}\text{Ar}/^{39}\text{Ar}$ geochronology of Middle to Late Cenozoic ignimbrites, mafic lavas, and volcanoclastic rocks in the Quemado Region, New Mexico. *Guidebook New Mexico Geol. Soc.* 45, 165–185 (1994).
23. WoldeGabriel, G. *et al.* Ecological and temporal context of early Pliocene hominids at Aramis, Ethiopia. *Nature* 371, 330–333 (1994).
24. Kirschvink, J. L. The least squares line and plane and the analysis of palaeomagnetic data. *R. Astron. Soc. Geophys. J.* 62, 699–718 (1980).
25. Fisher, R. A. Dispersion on a sphere. *Proc. R. Soc. Lond. A* 217, 295–305 (1953).

Supplementary Information accompanies the paper on www.nature.com/nature.

Acknowledgements The L.S.B. Leakey Foundation provided a major grant for the Gona Project, and the National Science Foundation (and Researching Hominid Origins Initiative-RHOI), the Wenner-Gren Foundation and the National Geographic Society funded the research. We thank N. Toth and K. Schick (co-directors of CRAFT Stone Age Institute, Indiana University) for their overall assistance. S.S. thanks N. Toth and K. Schick, and Friends of CRAFT for the Research Associate position at the Institute. The research permission by the Ministry of Youth, Sports and Culture, the Authority for Research and Conservation of Cultural Heritage and the National Museum of Ethiopia is greatly appreciated. We thank the Afar Regional Administration and the Afar colleagues at Eloha and Asayta for their hospitality, and A. Humet for the hard work in the field. Y. Beyene, C. Howell, B. Lasher and A. Almquist encouraged the research. D. Stout, L. Harlacker, M. Everett and T. Pickering assisted in the field, and A. Tamburro at CRAFT. The manuscript has benefited from discussions with B. Asfaw, M. Asnake, R. E. Bernor, J.-R. Boisserie, M. Brunet, S. Frost, Y. Haile-Selassie, O. Lovejoy, M. Pickford, E. Smith, H. Wesselman and T. White. M. Sahnouni and B. Smith helped with computer graphics.

Competing interests statement The authors declare that they have no competing financial interests.

Correspondence and requests for materials should be addressed to S.S. (ssemaw@indiana.edu).

Definitive fossil evidence for the extant avian radiation in the Cretaceous

Julia A. Clarke^{1,2}, Claudia P. Tambussi³, Jorge I. Noriega⁴, Gregory M. Erickson^{5,6,7} & Richard A. Ketcham⁸

¹Department of Marine, Earth and Atmospheric Sciences, North Carolina State University, Campus Box 8208, Raleigh, North Carolina 27695, USA

²North Carolina Museum of Natural Sciences, 11 West Jones Street, Raleigh, North Carolina 27601-1029, USA

³Museo de La Plata-CONICET, Paseo del Bosque s/n. La Plata (1900), Argentina

⁴Centro de Investigaciones Científicas y TTP- CONICET, Matteri y España, 3105 Diamante, Entre Rios, Argentina

⁵Department of Biological Science, Florida State University, Conradi Building, Dewey Street and Palmetto Drive, Tallahassee, Florida 32306-1100, USA

⁶Division of Paleontology, American Museum of Natural History, Central Park West at 79th Street, New York, New York 10024-5192, USA

⁷Department of Geology, The Field Museum, 1400 South Lake Shore Drive, Chicago, Illinois 60605, USA

⁸High-Resolution X-Ray Computed Tomography Facility, Jackson School of Geosciences, University of Texas at Austin, 1 University Station, C-1100, Austin, Texas 78712-0254, USA

Long-standing controversy^{1–9} surrounds the question of whether living bird lineages emerged after non-avian dinosaur extinction at the Cretaceous/Tertiary (K/T) boundary^{1,6} or whether these lineages coexisted with other dinosaurs and passed through this mass extinction event^{2–5,7–9}. Inferences from biogeography^{4,8} and molecular sequence data^{2,3,5,9} (but see ref. 10) project major avian lineages deep into the Cretaceous period, implying their ‘mass

survival’¹³ at the K/T boundary. By contrast, it has been argued that the fossil record refutes this hypothesis, placing a ‘big bang’ of avian radiation only after the end of the Cretaceous^{1,6}. However, other fossil data—fragmentary bones referred to extant bird lineages^{11–13}—have been considered inconclusive^{1,6,14}. These data have never been subjected to phylogenetic analysis. Here we identify a rare, partial skeleton from the Maastrichtian of Antarctica¹⁵ as the first Cretaceous fossil definitively placed within the extant bird radiation. Several phylogenetic analyses supported by independent histological data indicate that a new species, *Vegavis iaai*, is a part of Anseriformes (waterfowl) and is most closely related to Anatidae, which includes true ducks. A minimum of five divergences within Aves before the K/T boundary are inferred from the placement of *Vegavis*; at least duck, chicken and ratite bird relatives were coextant with non-avian dinosaurs.

The *Vegavis iaai* holotype specimen from Vega Island, western Antarctica, was discovered in 1992 and received rudimentary preparation that, in fact, degraded delicate bones that were originally exposed. It was reported¹⁵ as a possible ‘transitional’ form close to extant lineages¹⁵. For a decade since, the specimen’s exact systematic position and possible crown clade avian status have been debated^{6,14,16,17}. Significant new preparation, X-ray computed tomography (CT)¹⁸ and recovery of latex peels of the specimen before its original preparation reveal numerous, previously unknown bones and anatomical details. These new data, when included serially in three of the largest cladistic data sets considering Avialae¹⁹, Aves²⁰ and Anseriformes¹⁶, establish hierarchically nested character support for the placement of *Vegavis*.

Aves Linnaeus, 1758 (*sensu* Gauthier, 1986)

Neognathae Pycraft, 1900

Anseriformes Wagler, 1831

Anatoidea Leach, 1820 (*sensu* Livezey, 1997)

Vegavis iaai sp. nov.

Etymology. ‘Vegavis’ is for the holotype specimen’s Vega Island provenance; ‘avis’ is from the Latin for bird; and ‘iaai’ is for the Instituto Antártico Argentino (IAA) expedition that collected the specimen.

Holotype. MLP 93-I-3-1 (Museo de La Plata, Argentina), a disarticulated partial postcranial skeleton preserved in two halves of a concretion (Figs 1 and 2; see Supplementary Information for additional CT scan images, photographs, character data and measurements). Newly uncovered elements include five thoracic vertebrae, two cervical vertebrae, left scapula, right ulna, all pelvic bones, right and left fibulae and left? tarsometatarsal shaft. Previously reported elements¹⁵ include the complete right humerus, proximal left humerus, right coracoid, femora, left tibiotarsus, distal right radius, sacrum, distal left (right of ref. 15) tarsometatarsus, proximal right (left of ref. 15) tarsometatarsus and more than six dorsal ribs.

Locality. Cape Lamb, Vega Island, locality VEG9303 of the 1992/1993 IAA expedition¹⁵. Deposits are near-shore marine fine-grain sandstones²¹ from the Middle? to Upper Maastrichtian (~66–68 million years ago (Myr)) lithostratigraphic unit K3 of ref. 21 (see Supplementary Information for locality, horizon and dating details).

Diagnosis. *Vegavis* is unique among the surveyed taxa (Fig. 3) in that it has a low ridge on the medial edge of the proximal tibiotarsus that is proposed to be an autapomorphy of the new species (Fig. 2). The additional unique combination of characters from the phylogenetic analyses that differentiate *Vegavis* are given in the Methods. **Description.** *Vegavis* has heterocoelous cervical and thoracic vertebrae and 14–15 fused sacral vertebrae (Fig. 1). The apneumatic coracoid is penetrated by a supracoracoideus nerve foramen (Fig. 1). The blade of the scapula is slightly curved and narrow (Fig. 1). Its

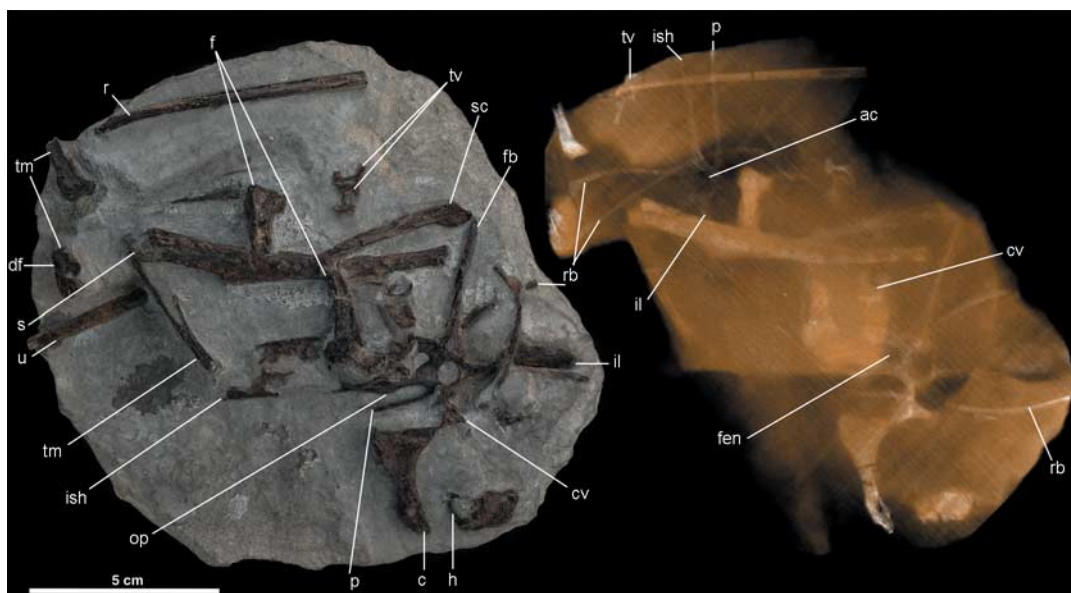


Figure 1 The half of the *Vegavis iaii* concretion that preserves most of MLP 93-I-3-1. Photograph (left) and volume renderings using CT data, highlighting the bone and rendering the matrix semi-transparent to elements preserved within the block (right). ac, acetabulum; c, coracoid; cv, cervical vertebra; df, distal vascular foramen; f, femora;

fb, fibula; fen, ilioischial fenestra; h, humerus; il, ilium; ish, ischium; op, obturator process; p, pubis; r, radius; rb, rib(s); tm, tarsometatarsus; tv, thoracic vertebrae; s, sacrum; sc, scapula; u, ulna.

coracoid tubercle is well projected and hemispherical. The humerus is slightly longer than the sacrum and about the length of a tibiotarsus (Fig. 2). Its deltopectoral crest is anteriorly deflected, less than shaft width and extends for approximately one-third of the shaft length (Fig. 2). A faint scar is developed in the location of the scapulohumeralis cranialis muscle insertion in Aves^{15,22} (Fig. 2). The capital ridge of the humeral shaft is strongly marked and the pneumotricipitalis fossa is shallow. The brachial scar angles obliquely, deepening ventrodistally into a fossa. The dorsodistal radius preserves one narrow ligamental groove.

The pelvic elements are firmly ankylosed to each other but may not have been fused to the sacrum. That the ilioischial fenestra was closed posteriorly is inferred from a flat sheet of bone (Fig. 1) preserved in both parts of the concretion. The postacetabular ilium is approximately twice the length of the preacetabular portion. A small pectineal process is present. The obturator foramen is elongate and posteriorly demarcated. The pubis is robust, straight, posteriorly directed and subparallel with the ischium (Fig. 1). The femur has a low trochanteric crest proximally (Fig. 1) and a patellar groove distally (Fig. 2). The proximal tibiotarsus preserves proximal portions of anterior and lateral cnemial crests (Fig. 2). The distal condyles are approximately the same width, and an ossified supratendinal bridge is developed over the extensor groove (Fig. 2). The diameter of the intercondylar groove is approximately one-third of the total distal tibiotarsal width. Metatarsals II–IV are fused throughout their length to enclose the distal vascular foramen (Fig. 1). Metatarsal II extends distally to approximately the base of metatarsal IV. There are four crests bounding three distinct hypotarsal sulci (Fig. 3c, insets). The medial hypotarsal crest is plantarly projected slightly farther than the other approximately equally projected crests.

The morphology of the *Vegavis* hypotarsus, with multiple, similarly proportioned canals, shares its derived structure with Anatidae (true ducks, geese and swans; Fig. 3c, insets). This feature, however, is only one of 20 unambiguously optimized synapomorphies preserved in *Vegavis* that support its placement as part of the interested clades Ornithurae, Aves, Neognathae, Anseriformes and Anatoidea, and finally, in an unresolved trichotomy with

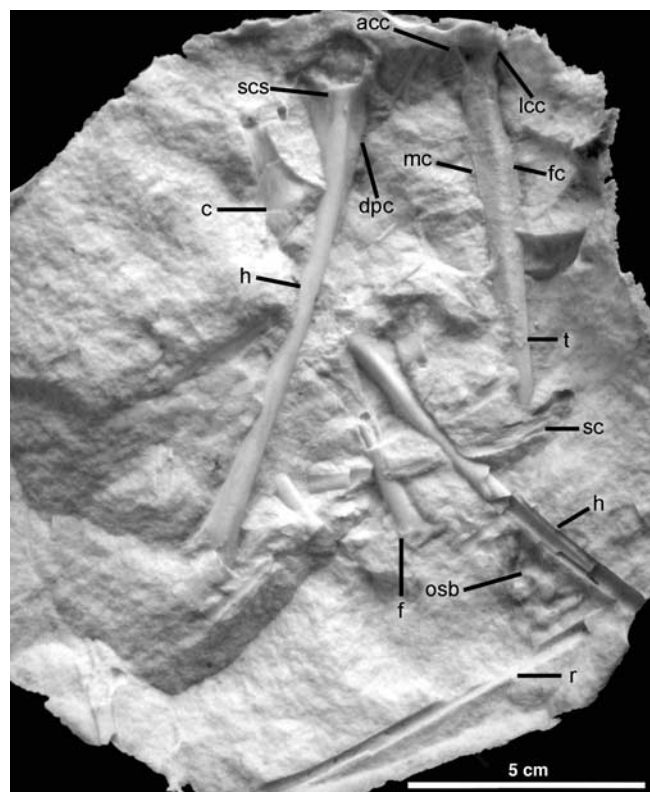


Figure 2 Recovered latex peel of the other half of the *Vegavis iaii* holotype block before original preparation. The coracoid, humerus and tibia were severely damaged when prepared out of this block. acc, anterior cnemial crest; c, coracoid; dpc, deltopectoral crest; f, femur; fc, fibular crest; h, humerus; lcc, lateral cnemial crest; mc, medial crest; osb, ossified supratendinal bridge; r, radius; t, tibiotarsus; sc, scapula; scs, scar of m. scapulohumeralis cranialis.

Presbyornis and Anatidae (Fig. 3). This placement of *Vegavis* implies a minimum of five cladogenetic events within Aves by the Upper Maastrichtian. The anseriform crown clade must be present (including three anseriform lineages with extant descendants) as well as parts of minimally stem-lineage neoavian neognaths, palaeognaths and galliforms.

Histological analysis of the *Vegavis* radius (using polarized microscopy) and examination of the humeral and femoral diaphyses (using dissecting microscopy) revealed most of the cortices to be composed of a highly vascularized (semi-reticular pattern) fibrolamellar matrix that grades into an avascular matrix periosteally²³. Lines of arrested growth (LAG or growth lines; Fig. 4) are absent in all specimens. Portions of the medullar cavities are lined by lamellar endosteal bone (Fig. 4). These characteristics suggest that the *Vegavis* holotype specimen was a somatically (skeletal) mature adult at the time of death²³. This suite of features is phylogenetically inconsistent with more common basal Mesozoic birds such as enantiornithines but supports placement of *Vegavis* within Ornithurae, a clade inclusive of extant bird lineages^{24,25} (Fig. 3). This conclusion is consistent with the independent phylogenetic results.

The placement of *Vegavis* confirms the origin of Aves and the

presence of several basal lineages by the latest Cretaceous. This result is compatible with either limited deep avian divergences by this time and, thus, limited survivorship at the K/T boundary^{9,10}, or the presence of most major lineages in the Cretaceous^{2-5,8} and 'mass survival'³ at this boundary. It contradicts a proposed early Tertiary crown clade origin^{1,6}; basal avian lineages were present with non-avian dinosaurs. Hypotheses implying a causal relationship between the extinction of non-avian dinosaurs and diversification of basal avian lineages^{9,26} must address these new data. *Vegavis* is the most complete Cretaceous specimen to be identified as part of the extant avian radiation and the first so identified through cladistic analyses; therefore, it provides the first reliable Cretaceous internal calibration point for 'molecular clock' approaches to dating the emergence of all living birds. It is strikingly close in age to some (for example, 66 Myr⁹) estimates of crown anseriform divergences made using these techniques^{9,10}. However, these estimates place most other major avian divergences earlier than or approximately contemporaneous with the Cretaceous^{2-5,9}, a proposal still unsupported by the fossil record in the Cretaceous. Only the lineage leading to the presently most speciose extant clade of birds, Neoaves, can be inferred present by the Maastrichtian from *Vegavis*' placement.

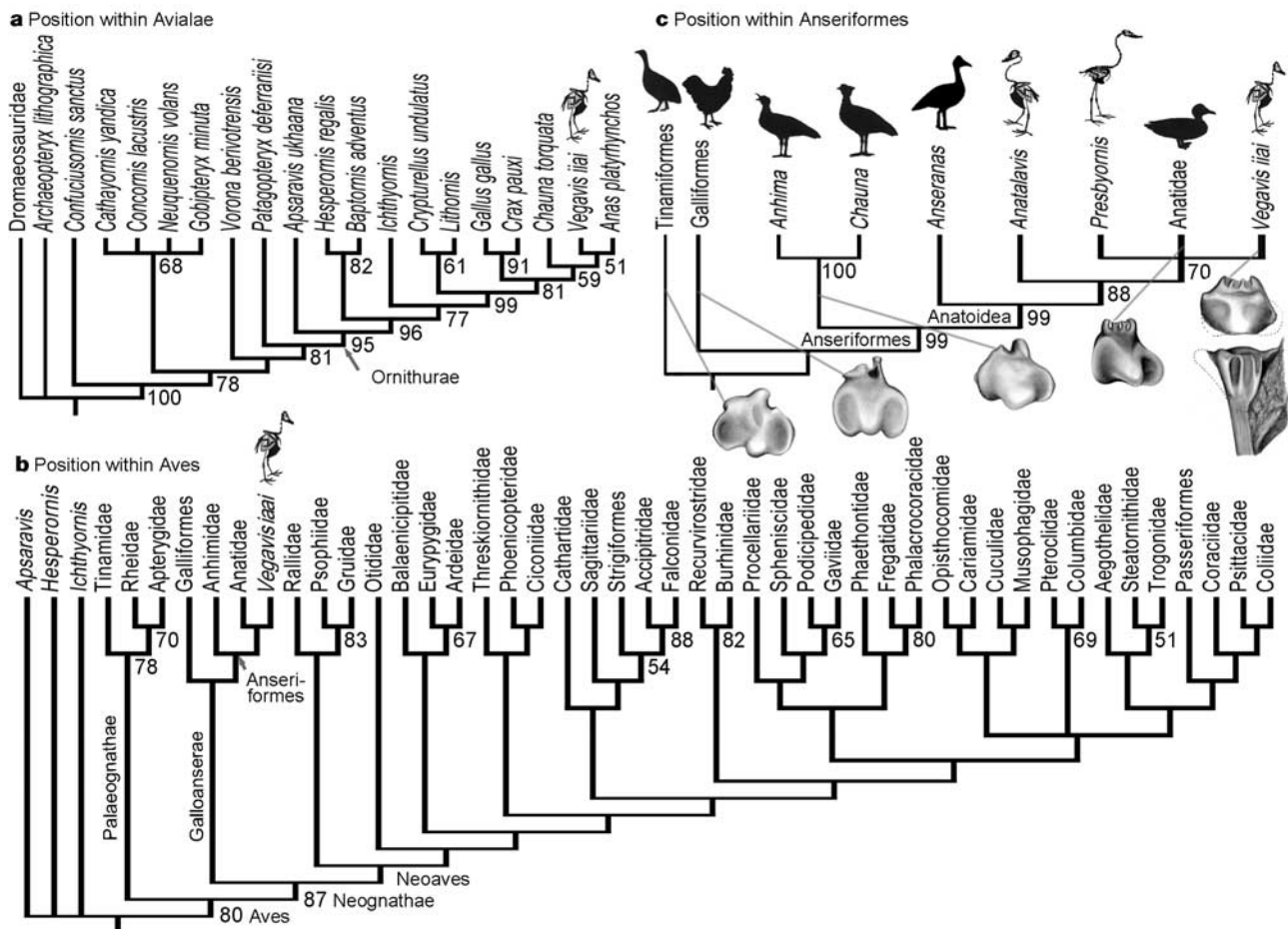


Figure 3 Phylogenetic placement of *Vegavis* in three successive cladistic analyses progressing from Avialae to Anseriformes (see Methods). **a**, Placement within Avialae in the strict consensus cladogram of two most parsimonious trees (MPTs): length, 385; consistency index (CI), 0.67; retention index (RI), 0.81; rescaled consistency index (RC), 0.54. **b**, Placement within Aves in the strict consensus of three MPTs: length, 822; CI, 0.33; RI, 0.48; RC, 0.16. **c**, Placement in Anseriformes in one MPT: length, 148; CI, 0.91; RI, 0.88; RC, 0.81. Bootstrap support values >50% from 2,000 replicates (10 random

addition sequences/replicate; random start trees; tree bisection reconnection) are reported below and to the right of corresponding nodes. Insets in **c** compare the right hypotarsus (see also character 90:0, ref. 16) of exemplars for Tinamiformes (*Eudromia elegans*), Galliformes (*Ortalis canicollis*), Anhimidae (*Chauna torquata*) and Anatidae (*Anas platyrhynchos*). All analyses used PAUP* 4.0b10 (ref. 28) and branches were collapsed if minimum length was 0. Character scoring of *Vegavis* in all data sets is given in the Supplementary Information.

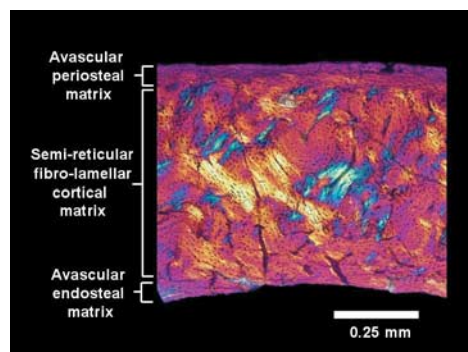


Figure 4 Histological section from the MLP 92-I-3-1 radius viewed with polarizing microscopy. The innermost seven-eighths of the cortex show a moderately vascularized, semi-reticular pattern, fibrolamellar matrix that is uninterrupted by lines of arrested growth. These data suggest that the animal showed relatively rapid, uninterrupted growth as in most living birds, including *Anatoidae*^{24,25}. The outermost cortices (top) are avascular indicating that a slowing of growth occurred, presumably as adulthood was approached. The presence of endosteal, avascular lamellar bone that partially lines the medullary cavity supports this developmental status interpretation. The primitive avialan long bone histological condition consists of moderately vascularized cortices with most vascular canals oriented longitudinally. The cortices in these birds are interrupted by lines of arrested growth (LAGs)^{24,25}. The vascular pattern and absence of LAGs in MLP 92-I-3-1 is consistent with its placement within *Ornithurae*²⁵ from the independent morphological character evidence.

Although the *Vegavis* holotype was originally suggested to be 'Presbyornithidae indeterminate'¹⁵, there is no evidence that it is part of this extinct taxon of predominately Eocene wading birds, averred by some to be transitional shorebird-duck 'mosaics'^{11,6,22,27}. Presbyornithids were proposed to be bridging taxa to waterfowl, indicating shorebird-like taxa to be the progenitors of all extant bird lineages^{1,6,22,27}. *Vegavis* has different proportions from *Presbyornis* that are closer to other extant basal anseriform species. Thus, there is further support¹⁶ that the wader proportions and the ecology used to diagnose Presbyornithidae^{17,22,27} are derived for that particular anseriform lineage and not ancestral avian characteristics. Finally, because of *Vegavis*' placement and its unknown skull morphology, advanced filter feeding cannot be assumed to be present in the anseriform lineage by the Maastrichtian. The Anseriformes that can be inferred as present by this point are lineages that today include large-bodied terrestrial browsers and occasional omnivores (that is, screamers, *Anhimidae* and magpie geese, *Anseranas*) as well as the lineage leading to true ducks and geese. □

Methods

Vegavis *iaai* was placed phylogenetically in three successive cladistic analyses progressing from Avialae to Anseriformes. Placement within Avialae was evaluated using the ref. 19 data set: 200 characters, 19 ingroup taxa; branch and bound search (Fig. 3a). Placement within Aves was evaluated using the ref. 20 data set: 148 characters, 46 ingroup taxa, heuristic search strategies of original publication (Fig. 3b). Placement in Anseriformes was evaluated using the ref. 16 data set: 123 characters, 8 ingroup taxa, branch and bound search (Fig. 3c). Extinct taxa *Anatalavis* and *Presbyornis* are included as the only other well-preserved basal anseriforms. *Anatalavis* is scored in this matrix from ref. 29.

The following unambiguously optimized synapomorphies are preserved in *Vegavis* and support its placement (character numbers in parentheses refer to the data sets referenced).

Ornithurae: at least 10 sacrales (61:4, ref. 19), domed humeral head (106:1, ref. 19), radius shaft with muscular impression (135:1, ref. 19), posterodorsal antitrochanter (158:1, ref. 19), pubis mediolaterally compressed (166:1, ref. 19), patellar groove present (172:1, ref. 19), distal tibiotarsal condyles equal in width (182:1, ref. 19) and proximal metatarsal III plantarly displaced (190:1, ref. 19).

Aves: anteriorly deflected humeral deltopectoral crest (112:1, ref. 19) that is less than shaft width (113:0, ref. 19), at least 15 ankylosed sacral vertebrae (61:6, ref. 19; 91:3, ref. 20) and ossified supratendinal bridge on tibiotarsus (100:1, ref. 20).

Neognathae: closed ilioischadic fenestra (94:1, ref. 20, 154:1, ref. 19) and humeral m. scapulotriceps groove (127:1, ref. 19; 81:1, ref. 20).

Anseriformes: diminutive pectineal process on pelvis (82:1, ref. 16) and hypotarsus with well developed cristae and sulci (103:12, ref. 20).

Anatoidae: lack of a sternal pneumatic foramen (70:0, ref. 16; apneumatic coracoid

90:0, ref. 19), ovoid m. scapulohumeralis cranialis scar (78:1, ref. 20) and metatarsal II shorter than IV (202:2, ref. 19). Lack of a pneumatic foramen on the proximomedial surface of ribs (59:2, ref. 16) and numerous hypotarsal cristae (90:0, ref. 16) are also synapomorphies of *Vegavis*, *Presbyornis* and Anatidae relative to *Anseranas*, but are ambiguously optimized because they are unknown in the Eocene *Anatalavis*. Loss of the n. supratoracoides foramen (65:1, ref. 20) and weak to absent thoracic vertebrae lateral excavations (58:0, ref. 19) are unambiguously optimized as synapomorphies of Anatidae relative to *Vegavis* and *Presbyornis*, but it is unresolved if *Vegavis* is the sister taxon of Anatidae, *Presbyornis*, or *Presbyornis* + Anatidae.

Received 16 July; accepted 19 October 2004; doi:10.1038/nature03150.

1. Feduccia, A. Explosive evolution in Tertiary birds and mammals. *Science* **267**, 637–638 (1995).
2. Hedges, S., Parker, P., Sibley, C. & Kumar, S. Continental breakup and the ordinal diversification of birds and mammals. *Nature* **381**, 226–229 (1996).
3. Cooper, A. & Penny, D. Mass survival of birds across the Cretaceous-Tertiary boundary: molecular evidence. *Science* **275**, 1109–1113 (1997).
4. Cracraft, J. Avian evolution, Gondwana biogeography and the Cretaceous-Tertiary mass extinction event. *Proc. R. Soc. Lond. B* **268**, 459–469 (2001).
5. van Tuinen, M. & Hedges, B. Calibration of avian molecular clocks. *Mol. Biol. Evol.* **18**, 206–213 (2001).
6. Feduccia, A. 'Big bang' for Tertiary birds? *Trends Ecol. Evol.* **18**, 172–176 (2003).
7. van Tuinen, M., Paton, T., Haddrath, O. & Baker, A. 'Big bang' for Tertiary birds? A replay. *Trends Ecol. Evol.* **18**, 442–443 (2003).
8. Ericson, P. G. P. et al. A Gondwanan origin of passerine birds supported by DNA sequences of the endemic New Zealand wrens. *Proc. R. Soc. Lond. B* **269**, 235–241 (2002).
9. Harrison, G. L. et al. Four new avian mitochondrial genomes help get to basic evolutionary questions in the late Cretaceous. *Mol. Biol. Evol.* **21**, 974–983 (2004).
10. Groth, J. & Barrowclough, G. Basal divergences in birds and the phylogenetic utility of the nuclear RAG-1 gene. *Mol. Phylogenet. Evol.* **12**, 115–123 (1999).
11. Stidham, T. A lower jaw from a Cretaceous parrot. *Nature* **396**, 29–30 (1998).
12. Kurochkin, E., Dyke, G. & Karhu, A. A new presbyornithid bird (Aves: Anseriformes) from the Late Cretaceous of southern Mongolia. *Am. Mus. Novit.* **3386**, 1–11 (2002).
13. Hope, S. in *Mesozoic Birds: Above the Heads of Dinosaurs* (eds Chiappe, L. M. & Witmer, L. M.) 339–388 (Univ. California Press, Berkeley, 2002).
14. Dyke, G. & van Tuinen, M. The evolutionary radiation of modern birds (Neornithes): reconciling molecules, morphology and the fossil record. *Zool. J. Linn. Soc.* **141**, 153–177 (2004).
15. Noriega, J. & Tambussi, C. A Late Cretaceous Presbyornithidae (Aves: Anseriformes) from Vega Island, Antarctic Peninsula: paleogeographic implications. *Ameghiniana* **32**, 57–61 (1995).
16. Livezey, B. The fossil *Presbyornis* and the interordinal relationships of waterfowl. *Zool. J. Linn. Soc.* **121**, 361–428 (1997).
17. Ericson, P. Systematic revision, skeletal anatomy, and paleoecology of the New World early Tertiary Presbyornithidae (Aves: Anseriformes). *PaleoBios* **20**, 1–23 (2000).
18. Ketcham, R. A. & Carlson, W. D. Acquisition, optimization and interpretation of X-ray computed tomographic imagery: Applications to the geosciences. *Comput. Geosci.* **27**, 381–400 (2001).
19. Clarke, J. & Norell, M. The morphology and phylogenetic position of *Apsaravis ukhaana* from the Late Cretaceous of Mongolia. *Am. Mus. Novit.* **3387**, 1–47 (2002).
20. Mayr, G. & Clarke, J. The deep divergences of neornithine birds: a phylogenetic analysis of morphological characters. *Cladistics* **19**, 553 (2003).
21. Marensi, S., Salani, S. & Santillana, S. Geología de cabo Lamb, isla Vega, Antártida. *Contribución Instituto Antártico Argentino* **530**, 1–43 (2001).
22. Ericson, P. New material of *Juncitarsus* (Phoenicopteridae), with a guide for differentiating that genus from the Presbyornithidae (Aves: Anseriformes). *Smithson. Contr. Paleobiol.* **89**, 245–251 (1999).
23. Francillon-Viellet et al. in *Biomimeticization: Patterns and Evolutionary Trends* (ed. Carter, J. G.) 471–530 (Van Nostrand Reinhold, New York, 1990).
24. Chinsamy, A. in *Mesozoic Birds: Above the Heads of Dinosaurs* (eds Chiappe, L. M. & Witmer, L. M.) 421–431 (Univ. California Press, Berkeley, 2002).
25. Padian, K., de Ricqlès, A. J. & Horner, J. R. Dinosaurian growth rates and bird origins. *Nature* **412**, 405–408 (2001).
26. Cooper, A. & Fortey, R. Evolutionary explosions and the phylogenetic fuse. *Trends Ecol. Evol.* **13**, 151–156 (1998).
27. Olson, S. & Feduccia, A. Relationships and evolution of flamingos (Aves: Phoenicopteridae). *Smithson. Contr. Zool.* **316**, 1–73 (1980).
28. Swofford, D. L. PAUP* Phylogenetic analysis using parsimony (*and other methods). Version 4.0. (Sinaur Associates, Sunderland, 2002).
29. Olson, S. The anseriform relationships of *Anatalavis* Olson and Parris (Anseranatidae), with a new species of the London Clay. *Smithson. Contr. Paleobiol.* **89**, 231–243 (1999).

Supplementary Information accompanies the paper on www.nature.com/nature.

Acknowledgements We thank Museo de La Plata for permission to CT scan and sample MLP 93-I-3-1 for histological analysis; M. Fox for reparation of the fossil; M. Reguero, S. Marensi and E. Olivero for stratigraphic information; T. Rowe and J. Humphries of UTCT lab for assistance with CT imaging; R. Edwards for photographs; A. Viñas for line drawings; B. Creisler for consultation on species name validity; and M. Norell for comments on the manuscript. Support for this project from an NSF Office of Polar Programs grant to J.A.C., the AMNH Division of Paleontology and Yale University is gratefully acknowledged.

Competing interests statement The authors declare that they have no competing financial interests.

Correspondence and requests for materials should be addressed to J.A.C. (Julia_Clarke@NCSU.edu).

Field parameterization and experimental test of the neutral theory of biodiversity

J. Timothy Wootton

Department of Ecology and Evolution, The University of Chicago, 1101 East 57th Street, Chicago, Illinois 60637, USA

Ecologists would like to explain general patterns observed across multi-species communities, such as species–area¹ and abundance–frequency relationships^{1–4}, in terms of the fundamental processes of birth, death and migration underlying the dynamics of all constituent species. The unified neutral theory of biodiversity^{5,6} and related theories^{7,8} based on these fundamental population processes have successfully recreated general species–abundance patterns without accounting for either the variation among species and individuals or resource-releasing processes such as predation and disturbance, long emphasized in ecological theory^{9–14}. If ecological communities can be described adequately without estimating variation in species and their interactions¹⁵, our understanding of ecological community organization and the predicted consequences of reduced biodiversity and environmental change would shift markedly. Here, I introduce a strong method to test the neutral theory that combines field parameterization of the underlying population dynamics with a field experiment, and apply it to a rocky intertidal community. Although the observed abundance–frequency distribution of the system follows that predicted by the neutral theory, the neutral theory predicts poorly the field experimental results, indicating an essential role for variation in species interactions.

Hubbell's neutral theory of biodiversity^{5,6} (Box 1) views individuals in ecosystems as being partitioned into a series of communities of fixed size at a local scale, linked by migration from a regional pool of individuals of fixed population size. By fixing the size of the local and regional communities, the theory assumes that limiting resource levels are constant and are used in their entirety by the individuals in the system. The individuals in the regional pool are distributed across a set of ecologically equivalent species, and the pool is influenced minimally by any single local community. Random speciation events increase the regional species pool, and species extinction may occur both locally and regionally through a random-walk process that depends on species' population size. Because of the neutrality assumption, the regional pool converges to a defined distribution of individuals among species, controlled by a function, known as the fundamental biodiversity number θ , which combines the regional population size and speciation rate. Local dynamics are typically scaled to a single death event and model the replacement of the dead individual as a function of the relative abundance of species in the regional pool, the relative species abundance in the local community, and probability of the new recruit being derived locally or via immigration from the regional pool.

As originally formulated, the neutral theory is accessible only to fairly weak empirical tests. Several of the key parameters underlying the model dynamics (regional population size, speciation rate, migration rate and death rate) are difficult or nearly impossible to measure directly. Hence, previous analyses have been limited to examining the consistency of the abundance–frequency distributions of local communities with distributions predicted by the neutral model. Such tests are weak, however, because both the observed and predicted data are ranked by abundance—guaranteeing that both observed and predicted functions decline monotonically—and because the predicted distributions are tuned by the θ

and migration rate parameters, permitting the predicted distributions to match a wide range of functional shapes¹⁶. A stronger test would be to generate estimates of model parameters from system dynamics, then test the predictions of the parameterized model in an independent situation that varied in known ways¹⁷, such as in a field experiment.

Box 1

Original neutral model and an empirically accessible, resource-based reformulation.

The neutral theory^{5,6} models the population dynamics at both local and regional scales, linked by recruitment from regional to local populations. The theory assumes ecologically identical individuals and complete utilization of limiting resources, resulting in approximately constant total numbers of individuals at both local and regional scales. Given these assumptions, the change in the size of a species' population at a local scale, modelled on the timescale of a single individual death, is

$$P_r(N_i - 1|N_i) = (N_i/J)[m(1 - p_i) + (1 - m)(\{J - N_i\}/\{J - 1\})]$$

$$P_r(N_i|N_i) = (N_i/J)[p_i m + (1 - m)(\{N_i - 1\}/\{J - 1\})] + (\{J - N_i\}/J) \times [m(1 - p_i) + (1 - m)(\{J - N_i - 1\}/\{J - 1\})] \quad (1)$$

$$P_r(N_i + 1|N_i) = (\{J - N_i\}/J)[p_i m + (1 - m)(N_i/\{J - 1\})]$$

where N_i is the local population size of species i , J is the total local population size, m is the probability that the individual replacing the dead individual is derived from the regional, rather than local, species pool, and p_i is the proportion of the regional species pool comprised of species i . The theory models the dynamics of the regional pool similarly, but random speciation events, rather than migration from an external source, add new species. The regional pool will converge to a known species abundance distribution, with a shape that depends on a parameter θ (referred to as the fundamental biodiversity number), which is a function of the total number of individuals contained in the regional pool and the per-individual speciation rate^{5,6}.

The local dynamical equations of the neutral theory can be reformulated by changing perspective from individuals in the community to units of limiting resource used by individuals. Empirically, assigning transitions in resource (for example, space) users is often easier than determining which particular individual replaced an individual that just died, and facilitates relaxing the assumption of saturated resource use. A resource-based reformulation of the local dynamics of the neutral theory is

$$P_r(i_t|i_0) = (1 - D) + D[p_i m + (L - m)N_{i,0}/J] \quad (2)$$

$$P_r(j_t|i_0) = D[p_j m + (L - m)N_{j,0}/J]$$

$$P_r(i_t|0) = p_i m + (L - m)N_{i,0}/J$$

where $P_r(j_t|i_0)$ is the probability that a unit of resource used by an individual of species i at an initial observation is used by species j at a subsequent observation t time units later, $P_r(i_t|0)$ is the probability that a unit of unused resource is subsequently used by an individual of species i , D is the probability that an individual using a resource unit does not survive over time t , and $N_{i,0}$ is the population size of species i in the local community at the initial observation point. A new parameter L ($m \leq L \leq 1$) allows exploration of recruitment limitation. When $L = 1$, no recruitment limitation to the resource occurs, as the neutral model assumes; $L < 1$ implies recruitment limitation.

Combining equation (2) and expanding across all resource units describes the dynamics of the entire local population of each species i

$$N_{i,t} = (1 - D)N_{i,0} + D[N_{i,0}(p_i m + (L - m)N_{i,0}/J) + (J - N_{i,0} - N_{0,0}) \times (p_i m + (L - m)N_{i,0}/J)] + N_{0,0}[p_i m + (L - m)N_{i,0}/J] \quad (3)$$

where $N_{0,0}$ is the amount of unused resource in the system. Parameters in these equations can be estimated with empirical data on species abundances and transition probabilities in resource use without knowing a priori the individual death rate, assuming that $L = 1$, or estimating θ .

By reformulating the theory to focus on resource units, particularly space or space-based resources, rather than individuals, many of the difficulties of parameter estimation can be overcome, permitting a strong test of the theory (Box 1). Specifically, by examining patterns of transitions among species using a known quantity of resource, death rate, migration rate and regional species composition can be estimated from local community dynamics, allowing parameterization of the local model without estimating θ , and allowing experimental tests of the model in systems where limiting resources can be readily identified. Additionally, this perspective permits evaluation of the consequences of relaxing the assumption of saturated resource use (Box 1).

Hard-substrate marine benthic communities are ideal for testing the neutral theory for several reasons. First, they are characterized by an easily quantified limiting resource: attachment space on rocks^{17–19}. Second, they exhibit relatively rapid dynamics, which allows relevant data on population dynamics to be collected over a reasonable timescale. Third, because of the moderate time and size scales over which the resident species operate, these systems have proven to be very amenable to field experimentation. Fourth, as predicted by the neutral model, they are well characterized as local communities linked to a regional source of individuals via oceanic dispersal. Finally, although studies in these systems have traditionally emphasized variation in species interactions^{13,18}, the distribution of abundance across species follows that predicted by the neutral model (Fig. 1), suggesting a possible paradox.

I assigned parameters to the local dynamics equations of the neutral theory using 15,200 transitions from a 12-yr record of species occupancy of points in space on intertidal rock benches on Tatoosh Island, Washington, USA (see Methods). Given the sampling design, a local community is defined in this study as the 100 individuals occupying points in a spatially fixed 60 × 60 cm quadrat. The transitions (Fig. 1) were fit to the reformulated neutral model (Box 1) with maximum likelihood methods to generate the best estimates of parameters at the local population scale, using the species proportions observed in independent transects (Fig. 1) to guide the search of parameter space (see Methods). Maximum likelihood parameterization of the transition data in unmanipulated plots generated estimates for death rate (D) and migration rate (m) of 0.25 and 0.68, respectively, which are biologically reasonable given the dispersal syndromes and lifespans typical of the sessile

species in this system. The migration rate estimates generated independently from the analysis of transition data and from fitting the model to the abundance–frequency distribution of the community (0.71, Fig. 1) are quite similar, further suggesting that the neutral model might apply.

To test the predictions of the parameterized neutral model, I performed a parallel experiment in which I chronically deleted the dominant species in the system, *Mytilus californianus*, from nine 60 × 60 cm plots, by hand-removing any individuals when they entered the plots over the 7–11-yr duration of the experiment (see Methods). By preventing the establishment of *M. californianus* into plots, I experimentally removed this species from both the local and regional pool, thereby shifting the regional pool composition parameters in a known way by proportionally increasing the remaining species (see Methods). I then compared the predicted abundance of the remaining species under the assumption of neutral dynamics with the results of the species deletion experiment.

Fitting the general shape of the abundance–frequency distribution to data derived from experimental plots again was consistent with the neutral model, suggesting that it might be applicable (Fig. 2). Because the underlying model was parameterized, however, I could make specific predictions about which species should be associated with each abundance value. When I compared experimental results to these more refined predictions, the neutral model was not supported (Fig. 3). The neutral model predictions differed significantly compared with the range of species compositions observed in the experimental plots (randomization tests of multivariate distances, $P < 0.001$, see Methods), and explained –29.6% of the variance in mean species abundance. Hence, the model predictions actually added error over an assumption that there was no variation in species abundance. One possible explanation for the poor fit of the neutral model might be the neutral model's assumption that all resources are used. I explored this possibility by using the reformulated model (Box 1) and introducing a recruitment limitation parameter (L) that was free to vary. This analysis yielded a best-fitting model with $m = 0.66$, $D = 0.10$ and $L = 0.66$. The equivalence of L and m implies that the local population contributed minimally to recruitment, which would not be surprising given the life histories of the sessile species in this system relative to the spatial scale of the local population. The predictions with these parameters, however, performed no better than the original neutral model ($P < 0.001$), explaining –30.3% of the variance in

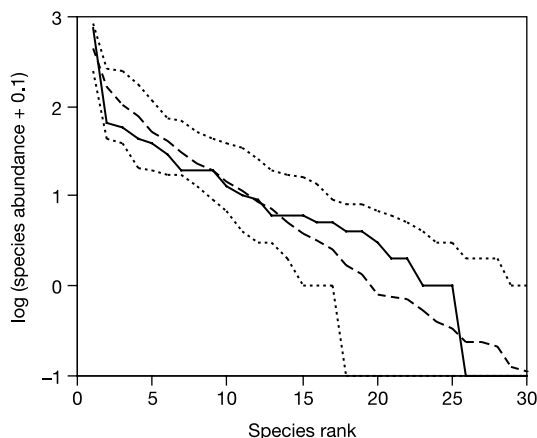


Figure 1 Comparison of species abundance distribution from randomly placed transects in the middle intertidal zone of Tatoosh Island, Washington, USA, compared to the predicted abundance distribution under the neutral model (with 95% confidence interval). Observed (solid line), mean predicted (dashed line) and 95% confidence interval predicted (dotted lines) curves are shown. Best estimates of parameters: $\theta = 5.7$, $m = 0.71$, $J = 1,128$ (parameter definitions in Box 1). Each species is arrayed along the x axis, ranked by abundance, and its abundance, on a log scale, is presented along the y axis.

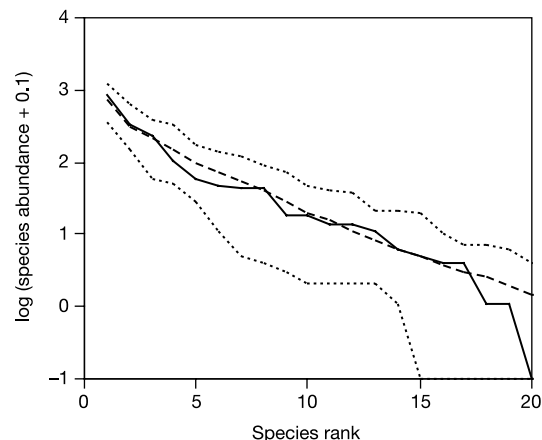


Figure 2 Ranked abundance frequency distribution of individuals from last two censuses in experimental plots compared to the predicted abundance distribution under the neutral model (with 95% confidence interval) made without reference to species identity. Observed (solid line), mean predicted (dashed line) and 95% confidence interval predicted (dotted lines) curves are shown. Best estimates of parameters: $\theta = 5.0$, $m = 0.80$, $J = 1,800$ (parameter definitions in Box 1).

mean species abundance in the experimental plots. Hence, the assumption of resource saturation could not explain the poor fit.

My results demonstrate that static abundance distribution patterns are a poor guide to identifying processes important to the organization of ecological communities, and illustrate the power of combining field experimentation with a focus on the population dynamic equations incorporating the essential processes of the neutral model, rather than on aggregate patterns several steps removed from ecological process. The results also clearly demonstrate that this community is not neutrally organized. The poor performance of the neutral model contrasts strongly with a Markov chain model incorporating species-specific differences²⁰, which successfully predicted 84% of the variation in mean species abundance from mussel removal experiments²¹. Parameter variation in the Markov chain model could arise from differences in local and regional population sizes of species, recruitment limitation, or differences in species traits and interspecific interactions. Because only the last factor was absent from the neutral model and extensions that I explored, species differences seem to be essential elements in generating the structure and dynamics of this intertidal community. Hence, accurate prediction of the system response to external impacts requires an accounting of variation in species interactions, and is not primarily determined by global abundance patterns. This conclusion does not preclude essential elements of the neutral model—including metacommunity structure, dispersal dynamics and speciation–extinction processes—from having important roles in community dynamics. For example, changes in the abundance patterns observed with increasing spatial scale of the local community^{4,6} might still be explained by sampling processes in a metacommunity context. Therefore, extending the theoretical framework of the neutral theory to include species differences may be rewarding.

If important assumptions of the neutral theory do not hold, why might patterns of species abundance follow a general pattern? One possibility is that species interaction strengths seem to follow a general pattern of distribution^{22–25}, which may constrain species abundance patterns. Initial theoretical work suggests that such interaction strength distributions may arise from community

assembly processes coupled with system stability and persistence^{26,27}. Investigations of these issues and links to macroecological pattern are still in their infancy, but appear profitable to pursue both empirically and theoretically. □

Methods

Details of data collection and experimental manipulation can be found elsewhere^{20,21}. Briefly, in 1993 I established 14 permanent 60 × 60 cm plots with 100 fixed points spaced regularly in a grid, and recorded the species present at each point in each plot annually in the late spring 1993–2004. I also collected independent data to characterize more broadly the species composition of the community by placing random transects through the mid-intertidal zone and identifying the sessile species present at each 5-cm interval along the transects²⁰.

Following prior studies, I fit the neutral model to static species abundance distributions derived from the independent transect data²⁰ following the algorithm outlined by Hubbell⁶, which was implemented in a program written with MatLab software. Briefly, this algorithm assembled a neutral community of the size of the data sample (the local community size J) by randomly assigning individuals to a species based on their current abundance in the community, the probability of a new recruit migrating from the regional pool (m), and the probability that a migrant is a new species, which is governed by the fundamental biodiversity number θ . I generated abundance distributions for a wide range of parameter values (0–1 for m , 1–20 for θ) and selected the parameter set that had the best fit to the distribution derived from the data, based on the maximum likelihood estimator⁶

$$L_e = \sum_{i=1}^s (N_i (\ln(N'_i/N_i) - (N'_i - N_i)))$$

where N'_i is the predicted abundance of species i for the sample size of the transect data, N_i is the actual abundance of species i in the transect data, and s is the number of species across all communities.

The local dynamic equations of the neutral model were fit using MatLab software to data from the annually censused plots using transitions in species occupancy at each point from one year to the next. One plot was not located in one year, reducing the number of transitions available by 200. The neutral theory was parameterized using these data by systematically evaluating parameter space for p_b , m , D and L in equation (3) (Box 1) and comparing the model fit to the data across all 14 plots and all 11 transition years using the maximum likelihood estimator

$$L_e = \sum_{t=1}^T \sum_{i=1}^s \sum_{j=1}^J (N_{i,j,t} (\ln(N'_{i,j,t}/N_{i,j,t}) - (N'_{i,j,t} - N_{i,j,t})))$$

where $N'_{i,j,t}$ is the predicted abundance of species i in local plot j at time t , and $N_{i,j,t}$ is the actual abundance of species i in plot j at time t . I estimated proportional species abundance of the regional pool (p_i) using the neutral theory to generate 1,000 distributions of species abundance, basing these distributions on the best estimate of θ derived from the species abundance distribution of 1,128 individuals from random transect data²⁰ (Fig. 1), and assuming a regional population of 10,000 individuals. Convergence of runs of the neutral model with 10,000 individuals and with larger population sizes is generally very high⁶. To maximize search efficiency, I assigned species to particular abundances generated by the neutral theory based on their abundance rank in the transect data²⁰, because the proportional species abundance of a large sample of individuals reflects the regional pool under the neutral model⁶. I also used the proportional abundance of species in the transect data as one possible estimate of the regional distribution. For each distribution of proportional abundance, I evaluated model fit while systematically varying m , D and L in 0.01-unit increments across the range of possible parameter values (0–1). The parameter combination that maximized the likelihood estimator was considered the proper description of neutral model dynamics. When zeros occurred in data or model predictions, I added 0.5 to both observed and predicted abundance to avoid problems of taking logs of zero.

I predicted the community response to deletion of *M. californianus* under the neutral model by setting $p_{Mc} = 0$ and adjusting the rest of the regional pool proportionately to sum to 1 (that is, $p'_i = p_i/(1 - p_{Mc})$, where p'_i is the adjusted proportion of species i in the regional pool after *M. californianus* (subscript Mc) is removed). I then simulated the neutral model for experimental plots of 100 individuals using the observed initial conditions and the duration of each of the nine replicates. Five other replicates²⁰ were not used because only data for per cent cover, not number of individuals at defined census points, were available and because the neutral theory predictions depend on knowing local population sizes. Analyses including these replicates and assuming a population size of 100 gave similar results. Each plot was simulated ten times under the neutral model, and the mean composition of these simulations was taken as the model prediction for each year. I compared the composition of individuals from the final two years of each experimental replicate and model simulation using a multivariate randomization approach. Specifically, I asked whether the mean of the absolute (euclidean) distance in multivariate space (that is, $[\sum_{i=1}^2 (N_{i,1} - N_{i,2})^2]^{0.5}$, where $N_{i,1}$ and $N_{i,2}$ are the number of individuals of species i present in community 1 and 2, respectively) between model predictions and experimental results for each of the nine plots tended to be higher than the mean euclidean distance between nine randomly chosen pairs of experimental plots. Higher distances between model predictions and experimental results would indicate a poorer fit of the model to the data than expected by chance. This comparison was repeated 1,000 times, and the fraction of times that model predictions had lower distances than experimental pairs was taken as the probability that the model predictions did not differ from the observed data. Other metrics (variance predicted, Bray–Curtis distance) were also examined to verify the robustness of the analysis, and did not change the conclusions.

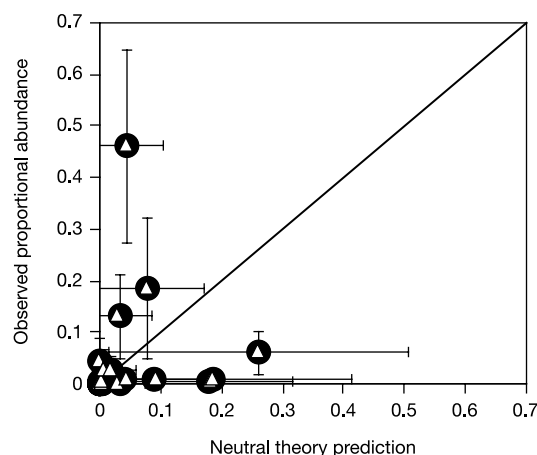


Figure 3 Comparison between species-specific predictions from the parameterized neutral model (filled circles) or model with recruitment limitation (open triangles) and observed proportional abundances in experimental mussel removal plots (error bars: 95% confidence intervals). Variations in mean predicted abundance in the neutral model arise from differences in species abundance generated by random-walk processes at both local and regional scales. Error bars for predictions arise from varying initial conditions in different experimental plots, and from stochastic effects on small local populations under multiple model realizations. The diagonal line is the expected relationship for perfect fit of experimental data to model predictions.

Received 24 September; accepted 16 November 2004; doi:10.1038/nature03211.

- MacArthur, R. H. & Wilson, E. O. *The Theory of Island Biogeography* (Princeton Univ. Press, Princeton, 1969).
- Fisher, R. A., Corbet, A. S. & Williams, C. B. The relation between the number of species and the number of individuals in a random sample of an animal population. *J. Anim. Ecol.* **12**, 42–58 (1943).
- Preston, F. W. The commonness, and rarity, of species. *Ecology* **41**, 611–627 (1948).
- Brown, J. H. *Macroecology* (Univ. Chicago Press, Chicago, 1995).
- Hubbell, S. P. A unified theory of biogeography and relative species abundance and its application to tropical rain forests and coral reefs. *Coral Reefs* **16**, S9–S21 (1997).
- Hubbell, S. P. *The Unified Theory of Biodiversity and Biogeography* (Princeton Univ. Press, Princeton, 2001).
- Caswell, H. Community structure: a neutral model analysis. *Ecol. Monogr.* **46**, 327–354 (1976).
- Bell, G. Neutral macroecology. *Science* **293**, 2413–2418 (2001).
- Elton, C. S. *Animal Ecology* (Sidgwick and Jackson, London, 1927).
- Gause, G. F. *The Struggle for Existence* (Hafner, New York, 1934).
- Hutchinson, G. E. Homage to Santa Rosalia or why are there so many kinds of animals? *Am. Nat.* **93**, 145–159 (1959).
- Huffaker, C. B. Experimental studies on predation: dispersion factors and predator-prey oscillations. *Hilgardia* **27**, 343–383 (1958).
- Paine, R. T. Food web complexity and species diversity. *Am. Nat.* **100**, 65–75 (1966).
- MacArthur, R. H. *Geographical Ecology* (Harper & Row, New York, 1972).
- Laska, M. S. & Wootton, J. T. Theoretical concepts and empirical approaches to measuring interaction strength. *Ecology* **79**, 461–476 (1998).
- McGill, B. J. Strong and weak tests of macroecological theory. *Oikos* **102**, 679–685 (2003).
- Adler, P. B. Neutral models fail to reproduce observed species-area and species-time relationships in Kansas grasslands. *Ecology* **85**, 1265–1272 (2004).
- Connell, J. H. The influence of interspecific competition and other factors on the distribution of the barnacle *Chthamalus stellatus*. *Ecology* **42**, 710–723 (1961).
- Paine, R. T. Ecological determinism in the competition for space. *Ecology* **65**, 1339–1348 (1984).
- Wootton, J. T. Prediction in complex communities: analysis of empirically-derived Markov models. *Ecology* **82**, 580–598 (2001).
- Wootton, J. T. Markov chain models predict the consequences of experimental extinctions. *Ecol. Lett.* **7**, 653–660 (2004).
- Paine, R. T. Food-web analysis through field measurements of per capita interaction strength. *Nature* **355**, 73–75 (1992).
- Moore, J. C., de Ruiter, P. C. & Hunt, H. W. The influence of productivity on the stability of real and model ecosystems. *Science* **261**, 906–908 (1993).
- Rafaelli, D. G. & Hall, S. J. In *Food Webs: Integration of Pattern and Dynamics* (eds Polis, G. & Winemiller, K.) 185–191 (Chapman and Hall, New York, 1996).
- Wootton, J. T. Estimates and tests of per-capita interaction strength: diet, abundance, and impact of intertidally foraging birds. *Ecol. Monogr.* **67**, 45–64 (1997).
- Kokkoris, G. D., Troumbis, A. Y. & Lawton, J. H. Patterns of species interaction strength in assembled theoretical competition communities. *Ecol. Lett.* **2**, 70–74 (1999).
- Drossel, B., McKane, A. & Quince, C. The impact of nonlinear functional responses on the long-term evolution of food web structure. *J. Theor. Biol.* **229**, 539–548 (2004).

Acknowledgements I thank the Makah Tribal Council for providing access to Tatoosh Island; J. Sheridan, J. Salamunovitch, F. Stevens, A. Miller, B. Scott, J. Chase, J. Shurin, K. Rose, L. Weis, R. Kordas, K. Edwards, M. Novak, J. Duke, J. Orcutt, K. Barnes, C. Neufeld and L. Weintraub for field assistance; and NSF, EPA (CISES) and the Andrew W. Mellon foundation for partial financial support.

Competing interests statement The author declares that he has no competing financial interests.

Correspondence and requests for materials should be addressed to J.T.W. (twoodton@uchicago.edu).

Evolutionary dynamics on graphs

Erez Lieberman^{1,2}, Christoph Hauert^{1,3} & Martin A. Nowak¹

¹Program for Evolutionary Dynamics, Departments of Organismic and Evolutionary Biology, Mathematics, and Applied Mathematics, Harvard University, Cambridge, Massachusetts 02138, USA

²Harvard-MIT Division of Health Sciences and Technology, Massachusetts Institute of Technology, Cambridge, Massachusetts, USA

³Department of Zoology, University of British Columbia, Vancouver, British Columbia V6T 1Z4, Canada

Evolutionary dynamics have been traditionally studied in the context of homogeneous or spatially extended populations^{1–4}. Here we generalize population structure by arranging individuals on a graph. Each vertex represents an individual. The weighted edges denote reproductive rates which govern how

often individuals place offspring into adjacent vertices. The homogeneous population, described by the Moran process⁵, is the special case of a fully connected graph with evenly weighted edges. Spatial structures are described by graphs where vertices are connected with their nearest neighbours. We also explore evolution on random and scale-free networks^{5–7}. We determine the fixation probability of mutants, and characterize those graphs for which fixation behaviour is identical to that of a homogeneous population⁷. Furthermore, some graphs act as suppressors and others as amplifiers of selection. It is even possible to find graphs that guarantee the fixation of any advantageous mutant. We also study frequency-dependent selection and show that the outcome of evolutionary games can depend entirely on the structure of the underlying graph. Evolutionary graph theory has many fascinating applications ranging from ecology to multi-cellular organization and economics.

Evolutionary dynamics act on populations. Neither genes, nor cells, nor individuals evolve; only populations evolve. In small populations, random drift dominates, whereas large populations

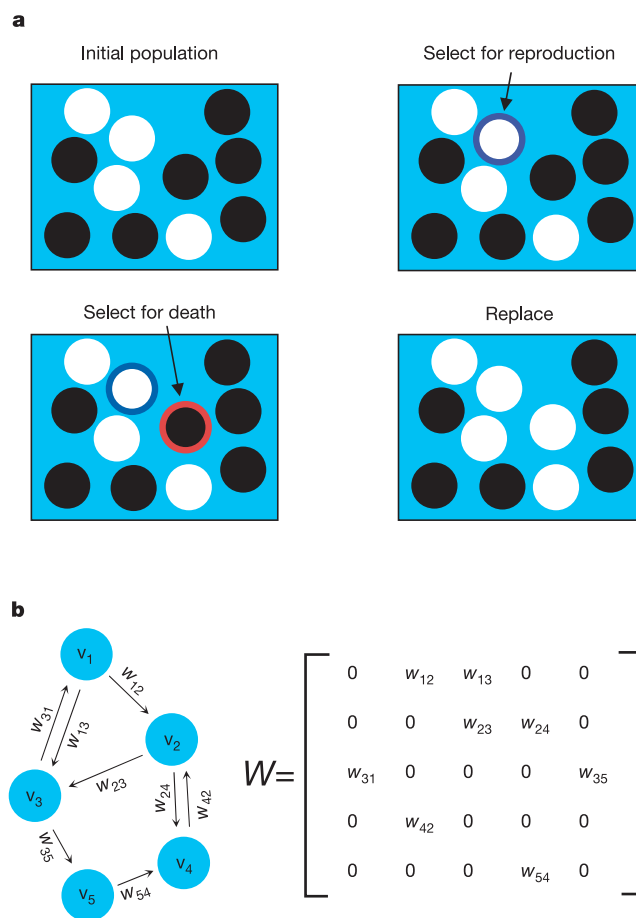


Figure 1 Models of evolution. **a**, The Moran process describes stochastic evolution of a finite population of constant size. In each time step, an individual is chosen for reproduction with a probability proportional to its fitness; a second individual is chosen for death. The offspring of the first individual replaces the second. **b**, In the setting of evolutionary graph theory, individuals occupy the vertices of a graph. In each time step, an individual is selected with a probability proportional to its fitness; the weights of the outgoing edges determine the probabilities that the corresponding neighbour will be replaced by the offspring. The process is described by a stochastic matrix W , where w_{ij} denotes the probability that an offspring of individual i will replace individual j . In a more general setting, at each time step, an edge ij is selected with a probability proportional to its weight and the fitness of the individual at its tail. The Moran process is the special case of a complete graph with identical weights.

are sensitive to subtle differences in selective values. The tension between selection and drift lies at the heart of the famous dispute between Fisher and Wright^{8–10}. There is evidence that population structure affects the interplay of these forces^{11–15}. But the celebrated results of Maruyama¹⁶ and Slatkin¹⁷ indicate that spatial structures are irrelevant for evolution under constant selection.

Here we introduce evolutionary graph theory, which suggests a promising new lead in the effort to provide a general account of how population structure affects evolutionary dynamics. We study the simplest possible question: what is the probability that a newly introduced mutant generates a lineage that takes over the whole population? This fixation probability determines the rate of evolution, which is the product of population size, mutation rate and fixation probability. The higher the correlation between the mutant's fitness and its probability of fixation, ρ , the stronger the effect of natural selection; if fixation is largely independent of fitness, drift dominates. We will show that some graphs are governed entirely by random drift, whereas others are immune to drift and are guided exclusively by natural selection.

Consider a homogeneous population of size N . At each time step an individual is chosen for reproduction with a probability proportional to its fitness. The offspring replaces a randomly chosen individual. In this so-called Moran process (Fig. 1a), the population size remains constant. Suppose all the resident individuals are identical and one new mutant is introduced. The new mutant has relative fitness r , as compared to the residents, whose fitness is 1. The fixation probability of the new mutant is:

$$\rho_1 = \frac{1 - 1/r}{1 - 1/r^N} \quad (1)$$

This represents a specific balance between selection and drift: advantageous mutations have a certain chance—but no guaran-

tee—of fixation, whereas disadvantageous mutants are likely—but again, no guarantee—to become extinct.

We introduce population structure as follows. Individuals are labelled $i = 1, 2, \dots, N$. The probability that individual i places its offspring into position j is given by w_{ij} .

Thus the individuals can be thought of as occupying the vertices of a graph. The matrix $W = [w_{ij}]$ determines the structure of the graph (Fig. 1b). If $w_{ij} = 0$ and $w_{ji} = 0$ then the vertices i and j are not connected. In each iteration, an individual i is chosen for reproduction with a probability proportional to its fitness. The resulting offspring will occupy vertex j with probability w_{ij} . Note that W is a stochastic matrix, which means that all its rows sum to one. We want to calculate the fixation probability ρ of a randomly placed mutant.

Imagine that the individuals are arranged on a spatial lattice that can be triangular, square, hexagonal or any similar tiling. For all such lattices ρ remains unchanged: it is equal to the ρ_1 obtained for the homogeneous population. In fact, it can be shown that if W is symmetric, $w_{ij} = w_{ji}$, then the fixation probability is always ρ_1 . The graphs in Fig. 2a–c, and all other symmetric, spatially extended models, have the same fixation probability as a homogeneous population^{17,18}.

There is an even wider class of graphs whose fixation probability is ρ_1 . Let $T_i = \sum_j w_{ji}$ be the temperature of vertex i . A vertex is 'hot' if it is replaced often and 'cold' if it is replaced rarely. The 'isothermal theorem' states that an evolutionary graph has fixation probability ρ_1 if and only if all vertices have the same temperature. Figure 2d gives an example of an isothermal graph where W is not symmetric. Isothermality is equivalent to the requirement that W is doubly stochastic, which means that each row and each column sums to one.

If a graph is not isothermal, the fixation probability is not given

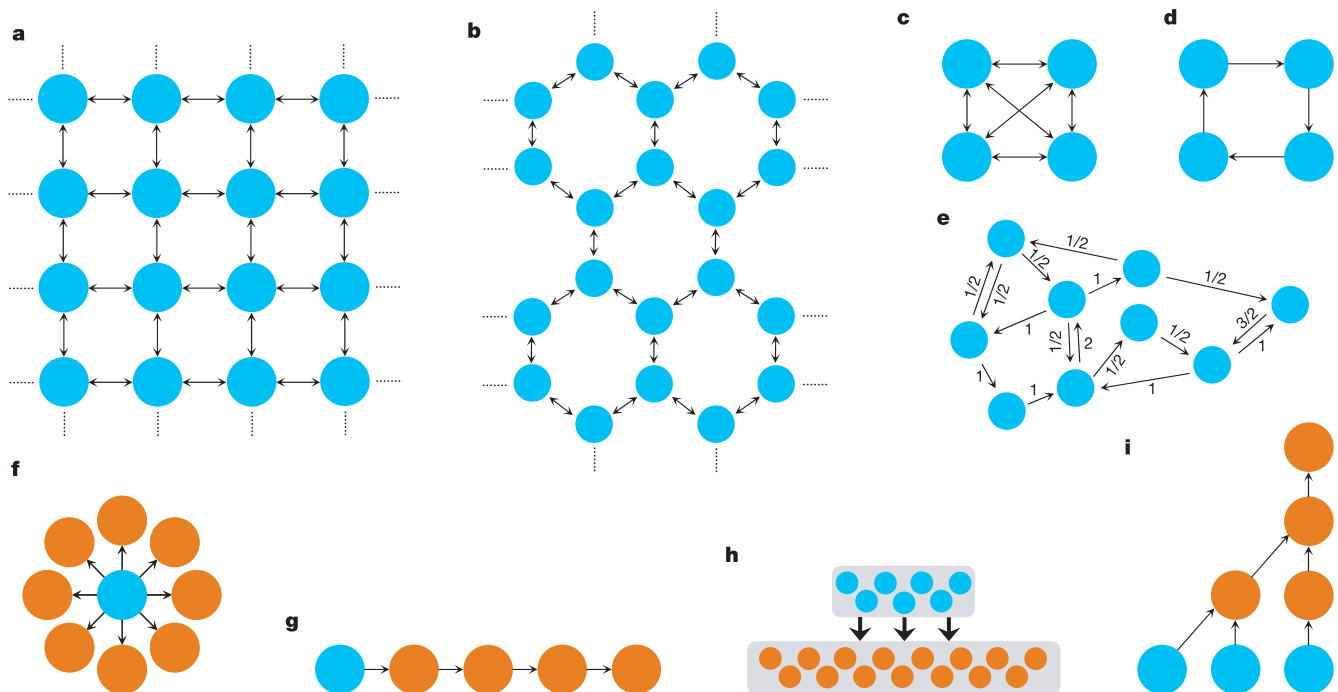


Figure 2 Isothermal graphs, and, more generally, circulations, have fixation behaviour identical to the Moran process. Examples of such graphs include: **a**, the square lattice; **b**, hexagonal lattice; **c**, complete graph; **d**, directed cycle; and **e**, a more irregular circulation. Whenever the weights of edges are not shown, a weight of one is distributed evenly across all those edges emerging from a given vertex. Graphs like **f**, the 'burst', and **g**, the 'path', suppress natural selection. The 'cold' upstream vertex is represented in

blue. The 'hot' downstream vertices, which change often, are coloured in orange. The type of the upstream root determines the fate of the entire graph. **h**, Small upstream populations with large downstream populations yield suppressors. **i**, In multirooted graphs, the roots compete indefinitely for the population. If a mutant arises in a root then neither fixation nor extinction is possible.

by ρ_1 . Instead, the balance between selection and drift tilts; now to one side, now to the other. Suppose N individuals are arranged in a linear array. Each individual places its offspring into the position immediately to its right. The leftmost individual is never replaced. What is the fixation probability of a randomly placed mutant with fitness r ? Clearly, it is $1/N$, irrespective of r . The mutant can only reach fixation if it arises in the leftmost position, which happens with probability $1/N$. This array is an example of a simple population structure whose behaviour is dominated by random drift.

More generally, an evolutionary graph has fixation probability $1/N$ for all r if and only if it is one-rooted (Fig. 2f, g). A one-rooted graph has a unique global source without incoming edges. If a graph has more than one root, then the probability of fixation is always zero: a mutant originating in one of the roots will generate a lineage which will never die out, but also never fixate (Fig. 2i). Small upstream populations feeding into large downstream populations are also suppressors of selection (Fig. 2h). Thus, it is easy to construct graphs that foster drift and suppress selection. Is it possible to suppress drift and amplify selection? Can we find structures where the fixation probability of advantageous mutants exceeds ρ_1 ?

The star structure (Fig. 3a) consists of a centre that is connected with each vertex on the periphery. All the peripheral vertices are connected only with the centre. For large N , the fixation probability

of a randomly placed mutant on the star is $\rho_2 = (1 - 1/r^2)/(1 - 1/r^{2N})$. Thus, any selective difference r is amplified to r^2 . The star acts as evolutionary amplifier, favouring advantageous mutants and inhibiting disadvantageous mutants. The balance tilts towards selection, and against drift.

The super-star, funnel and metafunnel (Fig. 3) have the amazing property that for large N , the fixation probability of any advantageous mutant converges to one, while the fixation probability of any disadvantageous mutant converges to zero. Hence, these population structures guarantee fixation of advantageous mutants however small their selective advantage. In general, we can prove that for sufficiently large population size N , a super-star of parameter K satisfies:

$$\rho_K = \frac{1 - 1/r^K}{1 - 1/r^{KN}} \quad (2)$$

Numerical simulations illustrating equation (2) are shown in Fig. 4a. Similar results hold for the funnel and metafunnel. Just as one-rooted structures entirely suppress the effects of selection, super-star structures function as arbitrarily strong amplifiers of selection and suppressors of random drift.

Scale-free networks, like the amplifier structures in Fig. 3, have most of their connectivity clustered in a few vertices. Such networks are potent selection amplifiers for mildly advantageous mutants (r

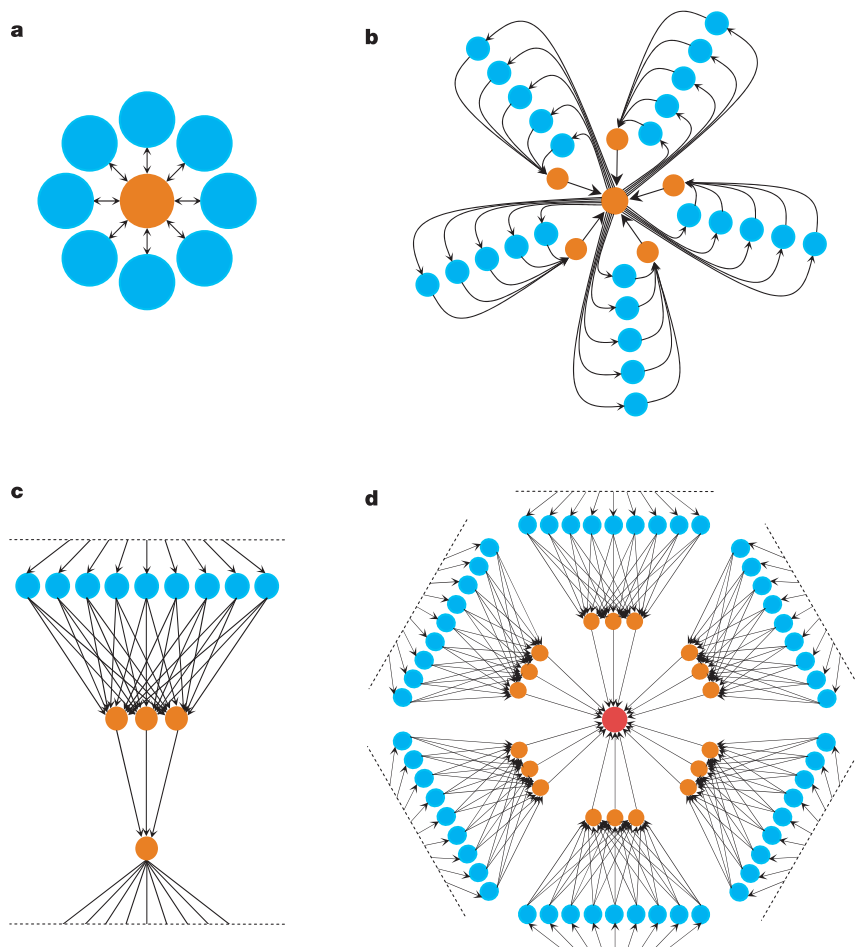


Figure 3 Selection amplifiers have remarkable symmetry properties. As the number of 'leaves' and the number of vertices in each leaf grows large, these amplifiers dramatically increase the apparent fitness of advantageous mutants: a mutant with fitness r on an amplifier of parameter K will fare as well as a mutant of fitness r^K in the Moran process. **a**, The star structure is a $K = 2$ amplifier. **b–d**, The super-star (**b**), the funnel (**c**) and the

metafunnel (**d**) can all be extended to arbitrarily large K , thereby guaranteeing the fixation of any advantageous mutant. The latter three structures are shown here for $K = 3$. The funnel has edges wrapping around from bottom to top. The metafunnel has outermost edges arising from the central vertex (only partially shown). The colours red, orange and blue indicate hot, warm and cold vertices.

close to 1), and relax to ρ_1 for very advantageous mutants ($r \gg 1$) (Fig. 4b).

Further generalizations of evolutionary graphs are possible. Suppose in each iteration an edge ij is chosen with a probability proportional to the product of its weight, w_{ij} , and the fitness of the individual i at its tail. In this case, the matrix W need not be stochastic; the weights can be any collection of non-negative real numbers.

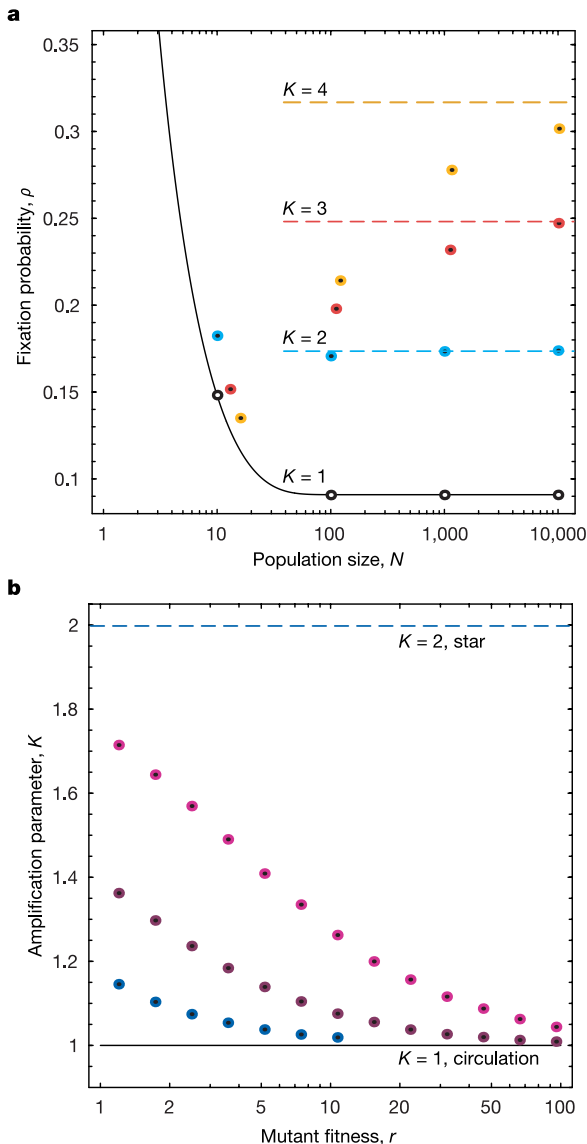


Figure 4 Simulation results showing the likelihood of mutant fixation. **a**, Fixation probabilities for an $r = 1.1$ mutant on a circulation (black), a star (blue), a $K = 3$ super-star (red), and a $K = 4$ super-star (yellow) for varying population sizes N . Simulation results are indicated by points. As expected, for large population sizes, the simulation results converge to the theoretical predictions (broken lines) obtained using equation (2). **b**, The amplification factor K of scale-free graphs with 100 vertices and an average connectivity of $2m$ with $m = 1$ (violet), $m = 2$ (purple), or $m = 3$ (navy) is compared to that for the star (blue line) and for circulations (black line). Increasing m increases the number of highly connected hubs. Scale-free graphs do not behave uniformly across the mutant spectrum: as the fitness r increases, the amplification factor relaxes from nearly 2 (the value for the star) to circulation-like values of unity. All simulations are based on 10^4 – 10^6 runs. Simulations can be explored online at <http://www.univie.ac.at/virtuallabs/>.

Here the results have a particularly elegant form. In the absence of upstream populations, if the sum of the weights of all edges leaving the vertex is the same for all vertices—meaning the fertility is independent of position—then the graph never suppresses selection. If the sum of the weights of all edges entering a vertex is the same for all vertices—meaning the mortality is independent of position—then the graph never suppresses drift. If both these conditions hold then the graph is called a circulation, and the structure favours neither selection nor drift. An evolutionary graph has fixation probability ρ_1 if and only if it is a circulation (see Fig. 2e). It is striking that the notion of a circulation, so common in deterministic contexts such as the study of flows, arises naturally in this stochastic evolutionary setting. The circulation criterion completely classifies all graph structures whose fixation behaviour is identical to that of the homogeneous population, and includes the subset of isothermal graphs (the mathematical details of these results are discussed in the Supplementary Information).

Let us now turn to evolutionary games on graphs^{18,19}. Consider, as before, two types A and B , but instead of having constant fitness, their relative fitness depends on the outcome of a game with payoff matrix:

$$\begin{array}{cc} & \begin{array}{c} A \quad B \end{array} \\ \begin{array}{c} A \\ B \end{array} & \begin{pmatrix} a & b \\ c & d \end{pmatrix} \end{array}$$

In traditional evolutionary game dynamics, a mutant strategy A can invade a resident B if $b > d$. For games on graphs, the crucial condition for A invading B , and hence the very notion of evolutionary stability, can be quite different.

As an illustration, imagine N players arranged on a directed cycle

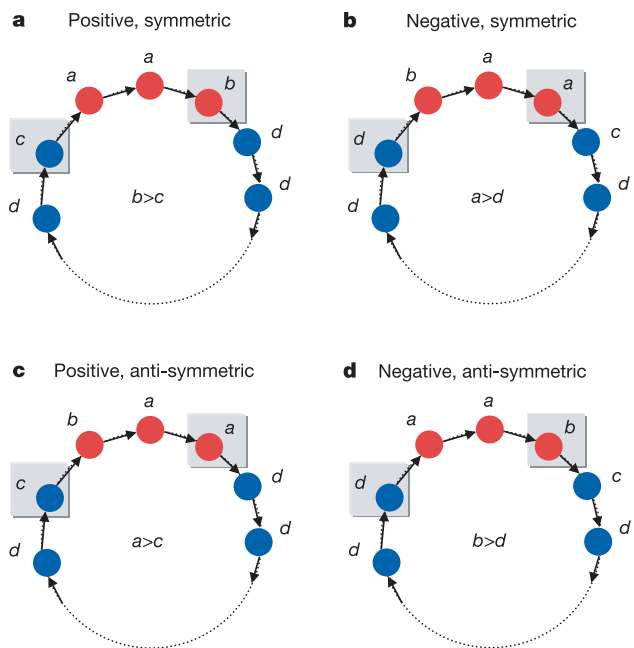


Figure 5 Evolutionary games on directed cycles for four different orientations. **a**, Positive symmetric. The invading mutant (red) is favoured over the resident (blue) if $b > c$. **b**, Negative symmetric. Invasion is favoured if $a > d$. For the Prisoner's Dilemma, the implication is that unconditional cooperators can invade and replace defectors starting from a single individual. **c**, Positive anti-symmetric. Invasion is favoured if $a > c$. The tables are turned: the invader behaves like a resident in a traditional setting. **d**, Negative anti-symmetric. Invasion is favoured if $b > d$. We recover the traditional invasion of evolutionary game theory.

(Fig. 5) with player i placing its offspring into $i + 1$. In the simplest case, the payoff of any individual comes from an interaction with one of its neighbours. There are four natural orientations. We discuss the fixation probability of a single A mutant for large N .

(1) Positive symmetric: i interacts with $i + 1$. The fixation probability is given by equation (1) with $r = b/c$. Selection favours the mutant if $b > c$.

(2) Negative symmetric: i interacts with $i - 1$. Selection favours the mutant if $a > d$. In the classical Prisoner's Dilemma, these dynamics favour unconditional cooperators invading defectors.

(3) Positive anti-symmetric: mutants at i interact with $i - 1$, but residents with $i + 1$. The mutant is favoured if $a > c$, behaving like a resident in the classical setting.

(4) Negative anti-symmetric: Mutants at i interact with $i + 1$, but residents with $i - 1$. The mutant is favoured if $b > d$, recovering the traditional invasion criterion.

Remarkably, games on directed cycles yield the complete range of pairwise conditions in determining whether selection favours the mutant or the resident.

Circulations no longer behave identically with respect to games. Outcomes depend on the graph, the game and the orientation. The vast array of cases constitutes a rich field for future study. Furthermore, we can prove that the general question of whether a population on a graph is vulnerable to invasion under frequency-dependent selection is NP (nondeterministic polynomial time)-hard.

The super-star possesses powerful amplifying properties in the case of games as well. For instance, in the positive symmetric orientation, the fixation probability for large N of a single A mutant is given by equation (1) with $r = (b/d)(b/c)^{K-1}$. For a super-star with large K , this r value diverges as long as $b > c$. Thus, even a dominated strategy ($a < c$ and $b < d$) satisfying $b > c$ will expand from a single mutant to conquer the entire super-star with a probability that can be made arbitrarily close to 1. The guaranteed fixation of this broad class of dominated strategies is a unique feature of evolutionary game theory on graphs: without structure, all dominated strategies die out. Similar results hold for the super-star in other orientations.

Evolutionary graph theory has many fascinating applications. Ecological habitats of species are neither regular spatial lattices nor simple two-dimensional surfaces, as is usually assumed^{20,21}, but contain locations that differ in their connectivity. In this respect, our results for scale-free graphs are very suggestive. Source and sink populations have the effect of suppressing selection, like one-rooted graphs^{22,23}.

Another application is somatic evolution within multicellular organisms. For example, the hematopoietic system constitutes an evolutionary graph with a suppressive hierarchical organization; stem cells produce precursors which generate differentiated cells²⁴. We expect tissues of long-lived multicellular organisms to be organized so as to suppress the somatic evolution that leads to cancer. Star structures can also be instantiated by populations of differentiating cells. For example, a stem cell in the centre generates differentiated cells, whose offspring either differentiate further, or revert back to stem cells. Such amplifiers of selection could be used in various developmental processes and also in the affinity maturation of the immune response.

Human organizations have complicated network structures^{25–27}. Evolutionary graph theory offers an appropriate tool to study selection on such networks. We can ask, for example, which networks are well suited to ensure the spread of favourable concepts. If a company is strictly one-rooted, then only those ideas that originate from the root will prevail (the CEO). A selection amplifier, like a star structure or a scale-free network, will enhance the spread of favourable ideas arising from any one individual. Notably, scientific collaboration graphs tend to be scale-free²⁸.

We have sketched the very beginnings of evolutionary graph

theory by studying the fixation probability of newly arising mutants. For constant selection, graphs can dramatically affect the balance between drift and selection. For frequency-dependent selection, graphs can redirect the process of selection itself.

Many more questions lie ahead. What is the maximum mutation rate compatible with adaptation on graphs? How does sexual reproduction affect evolution on graphs? What are the timescales associated with fixation, and how do they lead to coexistence in ecological settings^{29,30}? Furthermore, how does the graph itself change as a consequence of evolutionary dynamics³¹? Coupled with the present work, such studies will make increasingly clear the extent to which population structure affects the dynamics of evolution. □

Received 10 September; accepted 16 November 2004; doi:10.1038/nature03204.

1. Liggett, T. M. *Stochastic Interacting Systems: Contact, Voter and Exclusion Processes* (Springer, Berlin, 1999).
2. Durrett, R. & Levin, S. A. The importance of being discrete (and spatial). *Theor. Popul. Biol.* **46**, 363–394 (1994).
3. Moran, P. A. P. Random processes in genetics. *Proc. Camb. Phil. Soc.* **54**, 60–71 (1958).
4. Durrett, R. A. *Lecture Notes on Particle Systems & Percolation* (Wadsworth & Brooks/Cole Advanced Books & Software, Pacific Grove, 1988).
5. Erdős, P. & Rényi, A. On the evolution of random graphs. *Publ. Math. Inst. Hungarian Acad. Sci.* **5**, 17–61 (1960).
6. Barabási, A. & Albert, R. Emergence of scaling in random networks. *Science* **286**, 509–512 (1999).
7. Nagylaki, T. & Lucier, B. Numerical analysis of random drift in a cline. *Genetics* **94**, 497–517 (1980).
8. Wright, S. Evolution in Mendelian populations. *Genetics* **16**, 97–159 (1931).
9. Wright, S. The roles of mutation, inbreeding, crossbreeding and selection in evolution. *Proc. 6th Int. Congr. Genet.* **1**, 356–366 (1932).
10. Fisher, R. A. & Ford, E. B. The "Sewall Wright Effect". *Heredity* **4**, 117–119 (1950).
11. Barton, N. The probability of fixation of a favoured allele in a subdivided population. *Genet. Res.* **62**, 149–158 (1993).
12. Whitlock, M. Fixation probability and time in subdivided populations. *Genetics* **164**, 767–779 (2003).
13. Nowak, M. A. & May, R. M. The spatial dilemmas of evolution. *Int. J. Bifurcation Chaos* **3**, 35–78 (1993).
14. Hauert, C. & Doebeli, M. Spatial structure often inhibits the evolution of cooperation in the snowdrift game. *Nature* **428**, 643–646 (2004).
15. Hofbauer, J. & Sigmund, K. *Evolutionary Games and Population Dynamics* (Cambridge Univ. Press, Cambridge, 1998).
16. Maruyama, T. Effective number of alleles in a subdivided population. *Theor. Popul. Biol.* **1**, 273–306 (1970).
17. Slatkin, M. Fixation probabilities and fixation times in a subdivided population. *Evolution* **35**, 477–488 (1981).
18. Ebel, H. & Bornholdt, S. Coevolutionary games on networks. *Phys. Rev. E* **66**, 056118 (2002).
19. Abramson, G. & Kuperman, M. Social games in a social network. *Phys. Rev. E* **63**, 030901(R) (2001).
20. Tilman, D. & Kareiva, P. (eds) *Spatial Ecology: The Role of Space in Population Dynamics and Interspecific Interactions* (Monographs in Population Biology, Princeton Univ. Press, Princeton, 1997).
21. Neuhauser, C. Mathematical challenges in spatial ecology. *Not. AMS* **48**, 1304–1314 (2001).
22. Pulliam, H. R. Sources, sinks, and population regulation. *Am. Nat.* **132**, 652–661 (1988).
23. Hassell, M. P., Comins, H. N. & May, R. M. Species coexistence and self-organizing spatial dynamics. *Nature* **370**, 290–292 (1994).
24. Reya, T., Morrison, S. J., Clarke, M. & Weissman, I. L. Stem cells, cancer, and cancer stem cells. *Nature* **414**, 105–111 (2001).
25. Skyrms, B. & Pemanle, R. A dynamic model of social network formation. *Proc. Nat. Acad. Sci. USA* **97**, 9340–9346 (2000).
26. Jackson, M. O. & Watts, A. On the formation of interaction networks in social coordination games. *Games Econ. Behav.* **41**, 265–291 (2002).
27. Asavathiratham, C., Roy, S., Lesieutre, B. & Verghese, G. The influence model. *IEEE Control Syst. Mag.* **21**, 52–64 (2001).
28. Newman, M. E. J. The structure of scientific collaboration networks. *Proc. Natl Acad. Sci. USA* **98**, 404–409 (2001).
29. Boyd, S., Diaconis, P. & Xiao, L. Fastest mixing Markov chain on a graph. *SIAM Rev.* **46**, 667–689 (2004).
30. Nakamaru, M., Matsuda, H. & Iwasa, Y. The evolution of cooperation in a lattice-structured population. *J. Theor. Biol.* **184**, 65–81 (1997).
31. Bala, V. & Goyal, S. A noncooperative model of network formation. *Econometrica* **68**, 1181–1229 (2000).

Supplementary Information accompanies the paper on www.nature.com/nature.

Acknowledgements The Program for Evolutionary Dynamics is sponsored by J. Epstein. E.L. is supported by a National Defense Science and Engineering Graduate Fellowship. C.H. is grateful to the Swiss National Science Foundation. We are indebted to M. Brenner for many discussions.

Competing interests statement The authors declare that they have no competing financial interests.

Correspondence and requests for materials should be addressed to E.L. (erez@erez.com).

Protein kinase A signalling via CREB controls myogenesis induced by Wnt proteins

Alice E. Chen¹, David D. Ginty² & Chen-Ming Fan¹

¹Department of Embryology, Carnegie Institution of Washington, 115 West University Parkway, Baltimore, Maryland 21210, USA

²Department of Neuroscience, Howard Hughes Medical Institute, Johns Hopkins University School of Medicine, Baltimore, Maryland 21205, USA

Select members of the Wnt family of secreted glycoproteins have been implicated in inducing the myogenic determinant genes *Pax3*, *MyoD* and *Myf5* during mammalian embryogenesis^{1,2}, but the mechanism of induction has not been defined. We describe an unexpected role for protein kinase A (PKA) signalling via CREB in this induction. Using a combination of *in vitro* explant assays, mutant analysis and gene delivery into mouse embryos cultured *ex vivo*, we demonstrate that adenylyl cyclase signalling via PKA and its target transcription factor CREB are required for Wnt-directed myogenic gene expression. Wnt proteins can also stimulate CREB-mediated transcription, providing evidence for a Wnt signalling pathway involving PKA and CREB. Our findings raise the possibility that PKA/CREB signalling may also contribute to other Wnt-regulated processes in embryonic patterning, stem cell renewal and cancer³.

Mammalian trunk skeletal muscle is derived from the somites, which are mesodermal segments flanking the embryonic midline. Upon receiving inductive signals from the neural tube and surface ectoderm, myogenic precursors in the dorsal somite (the dermomyotome) first express *Pax3*. Subsequently, cells at the medial and lateral edges of the dermomyotome initiate expression of *Myf5* and *MyoD*, respectively. These cells delaminate from the dermomyotome to form the myotome. The medial myotome gives rise to the deep back muscles and the lateral myotome gives rise to the limb and body wall muscles⁴. *Pax3*, *Myf5* and *MyoD* are essential for initiation of the muscle programme^{5,6}. Wnt proteins such as Wnt1 (expressed in the dorsal neural tube) and Wnt7a (expressed in the lateral surface ectoderm) can induce these myogenic determinant genes *in vitro*^{1,2,7}. Although the canonical Wnt/ β -catenin pathway is typically implicated in cell fate-specific gene activation, neither an activated form of β -catenin⁸ nor LiCl treatment (not shown), which mimics β -catenin activation, can initiate myogenesis. Non-canonical Wnt pathways, identified based on homology between the Frizzled family of Wnt receptors and heterotrimeric G protein-coupled receptors⁹, use G protein effectors including protein kinase C, Ca²⁺ and cyclic GMP. However, these effectors exert effects primarily on cell morphology and behaviour^{10–12} rather than gene expression.

We established the *in vivo* relevance of adenylyl cyclase (AC) pathway activity in myogenesis by assaying for a downstream effector of this cascade, CREB, in embryonic regions undergoing myogenesis. CREB is the founding member of the CREB family of basic leucine zipper transcription factors (including ATF-1 and CREM), which bind the cyclic AMP-responsive element (CRE) as homodimers or heterodimers and activate transcription upon phosphorylation by PKA¹³. Embryonic regions where CREB is phosphorylated on its transcriptional regulatory site, Ser 133 (P-CREB), may represent potential areas of PKA activity. We performed immunohistochemistry on mouse sections at embryonic day 9.5 (E9.5) and E10.5 using antibodies against CREB and P-CREB. In contrast to the ubiquitous expression of CREB (see Supplementary Fig. S1), P-CREB was detected in myogenic regions of the somite (Fig. 1). At the time of somite formation, P-CREB

staining was found in the dorsal somitic cells (Fig. 1a), where *Pax3* expression was first upregulated (Fig. 1b). As somites are patterned along the dorso-ventral axis to form the dermomyotome and sclerotome, we found strong P-CREB (Fig. 1c) and *Pax3* (Fig. 1d) staining in the entire dermomyotome. As the myotome forms, high levels of P-CREB were found in cells at the medial and lateral dermomyotomal edges (Fig. 1e, f), where *Myf5* (Fig. 1h) and *MyoD* (Fig. 1g) expression was initiated. As cells move into the myotome proper, where *MyoD* and *Myf5* continue to be expressed, P-CREB was downregulated to a level slightly above that of the sclerotome (Fig. 1e). The level of P-CREB was also lower in the more mature dermomyotome compared with earlier stages (compare Fig. 1c and e). Thus, the dynamic patterns of P-CREB in the somite coincide with the patterns of myogenic gene induction.

These results led us to examine whether CREB null mutants¹⁴ (*CREB*^{−/−}) exhibit myogenic defects. *CREB*^{−/−} mice are born at reduced mendelian frequency¹⁴ (15%). We found that *CREB*^{−/−} mice were present at the expected 25% ratio between E9.5–E11.5,

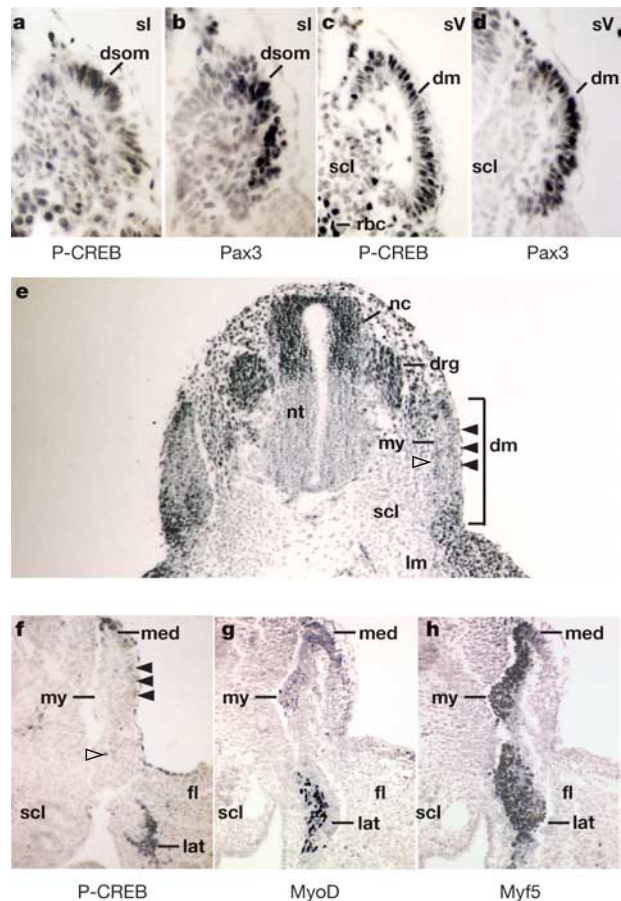


Figure 1 P-CREB staining patterns coincide with myogenic induction. Immunohistochemistry was performed on transverse sections of mouse embryos at E9.5 (a–d) and E10.5 (e–h). a, c, P-CREB and (b, d) *Pax3* staining in the first forming (sl) and fifth (sV) somites, dorsal somitic cells (dsom) and dermomyotome (dm). e–h, Staining at forelimb (fl) level. e, P-CREB correlates with regions patterned by Wnt proteins, for example, the dermomyotome edges, dorsal neural tube (nt), neural crest (nc), dorsal root ganglion (drg) and limb mesenchyme (lm). P-CREB (magnified in f) overlaps with *MyoD* expression (g) at the lateral (lat) edge and *Myf5* expression (h) at the medial (med) edge. Open and black arrowheads indicate reduced P-CREB levels in the myotome (my) and dermomyotome, respectively. Red blood cells (rbc) and a few sclerotome (scl) cells are P-CREB-positive (c), significance unknown.

but by E12.5 approximately half of them died, indicating that lethality occurs between E11.5 and E12.5. At E10.5, $CREB^{-/-}$ embryos existed as two populations, distinguishable by gross morphology: mild ' $CREB^{-/-}M$ ' (11%) and severe ' $CREB^{-/-}S$ ' (15%) mutants ($n = 88$). $CREB^{-/-}M$ embryos were comparable in size to wild-type littermates, whereas $CREB^{-/-}S$ embryos were ~35% smaller. In $CREB^{-/-}S$ somites, the rate of cell proliferation was decreased by ~30%, but no increase in the rate of apoptosis was observed (see Supplementary Fig. S2).

Whole-mount *in situ* hybridization (WISH) was employed to analyse wild-type, $CREB^{-/-}M$ and $CREB^{-/-}S$ embryos at E10.5 (embryos of similar size and stage were chosen for normalization of analyses, Fig. 2a–t). The domain of *Pax3* expression appeared smaller in the mutants compared with the wild type ($n = 3$; compare square brackets in Fig. 2b, d, f). Despite its reduced levels,

Pax3 expression was found in the dermomyotome of all somites ($CREB^{-/-}M$, $n = 3$; $CREB^{-/-}S$, $n = 3$). Strong expression of *MyoD* in tightly clustered cells was found in the lateral myotome of wild-type embryos ($n = 6$; Fig. 2h). Mutants had reduced and punctate *MyoD* expression, which was restricted to the more mature anterior somites ($CREB^{-/-}M$, $n = 5$; $CREB^{-/-}S$, $n = 4$; Fig. 2j, l). In wild-type embryos, the *Myf5* expression domain formed a well defined flat medial edge ($n = 3$; Fig. 2n). In the mutant embryos, *Myf5* expression at the medial edge appeared disorganized and pointed in shape ($CREB^{-/-}M$, $n = 3$; $CREB^{-/-}S$, $n = 3$; Fig. 2p, r). The myotome proper also had many fewer *Myf5*-positive cells. Interestingly, the *Pax3* and *Myf5* defects observed in the $CREB$ mutants are reminiscent of those found in $Wnt1^{-/-}$ $Wnt3a^{-/-}$ embryos¹⁵. We also noted that *Myf5* expression was not found at the cervical somites of $CREB$ mutants, suggesting an anterior-specific

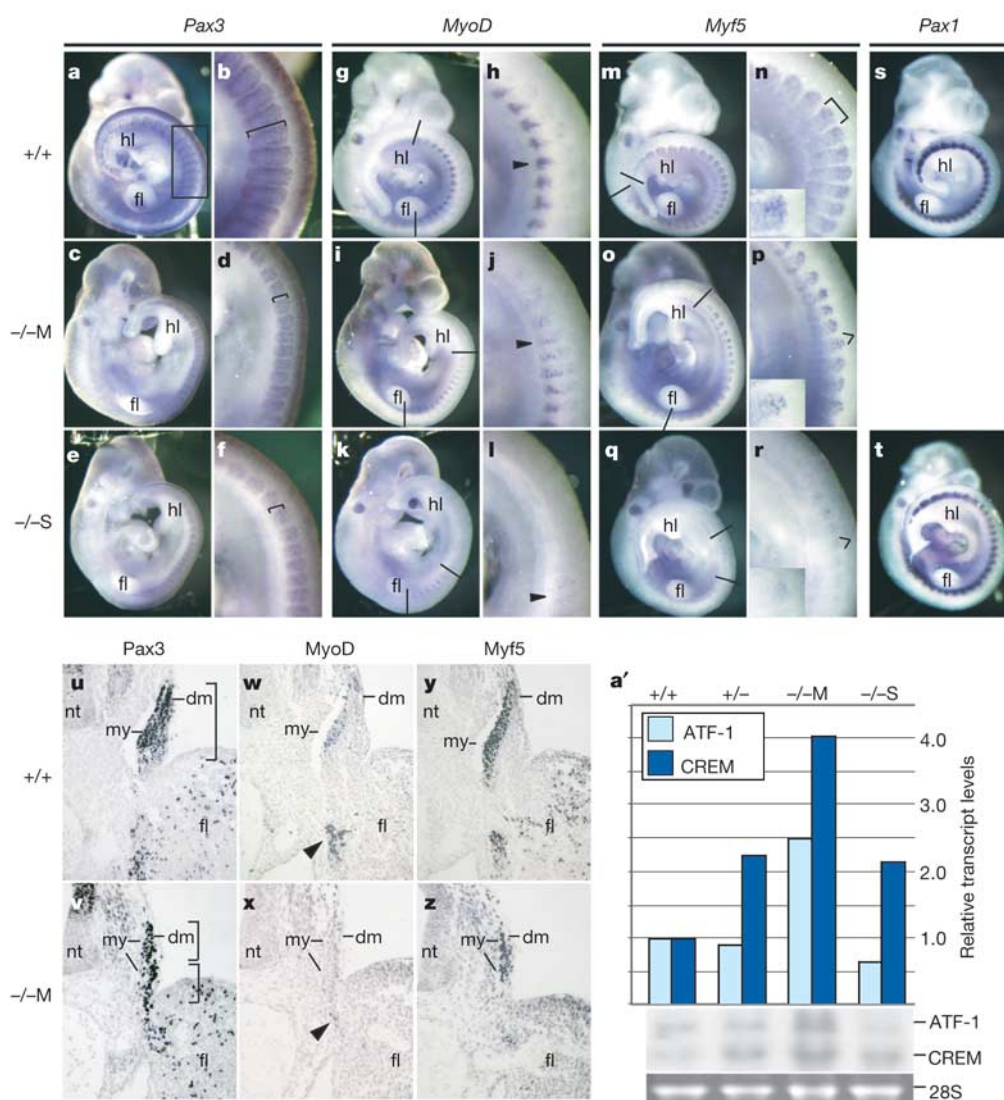


Figure 2 $CREB$ mutants have myogenic defects. **a–t**, WISH analysis of *Pax3* (**a–f**, trunk enlarged in **b**, **d**, **f**), *MyoD* (**g–l**, trunk enlarged in **h**, **j**, **l**), *Myf5* (**m–r**, trunk enlarged in **n**, **p**, **r**) and *Pax1* (**s**, **t**) in E10.5 wild-type (+/+) (top row), $CREB^{-/-}M$ (middle row) and $CREB^{-/-}S$ (bottom row) mice. Box in **a** shows the enlarged area. Square brackets (**b**, **d**, **f**) indicate *Pax3* domain size. Arrowheads (**h**, **j**, **l**) indicate *MyoD* expression. Square bracket (**n**) and angled brackets (**p**, **r**) compare the shape of *Myf5*-positive medial edges (enlarged in insets). Black lines indicate expression boundaries along the anterior-posterior axis, forelimb (fl), hindlimb (hl). **u–z**, Immunohistochemistry of *Pax3*, *MyoD* and *Myf5* on

adjacent transverse sections at the forelimb level of E10.5 wild-type (upper row) and $CREB^{-/-}M$ (lower row) mice. Brackets in **u**, **v** denote dermomyotome (dm) size, arrowheads in **w**, **x** show *MyoD*-positive cells; myotome (my), neural tube (nt). **a'**, Northern analysis using ATF-1 and CREM probes on total RNA of wild-type, $CREB^{+/-}$, $CREB^{-/-}M$ and $CREB^{-/-}S$ mice. 28S RNA reflects sample loads. Signals are shown as an X-ray film image (bottom); quantified and plotted as fold differences relative to wild-type levels (top).

requirement for CREB function. Importantly, the defects of *MyoD* and *Myf5* expression were more pronounced in $CREB^{-/-}$ S than in $CREB^{-/-}$ M embryos. These defects were not due to alterations in *Wnt1* or *Wnt7a*, as their expression was normal in CREB mutants (see Supplementary Fig. S3). Finally, we observed no obvious change in *Pax1* (a sclerotome marker) expression in the somite of $CREB^{-/-}$ S embryos ($n = 3$; compare Fig. 2s and t), indicating that the somitic defects are specific to the myogenic lineages.

These defects were substantiated by analysing *Pax3*, *MyoD* and *Myf5* expression by immunohistochemistry on sectioned E10.5 wild-type and CREB mutant embryos (Fig. 2u–z). At the forelimb level of $CREB^{-/-}$ M embryos, the size of the *Pax3*-positive domain was reduced. Only a few weakly *MyoD*-positive cells at the lateral edges were detected. *Myf5*-positive cells in the medial edge and myotome proper were reduced and disorganized. Even fewer *Pax3*-, *MyoD*- and *Myf5*-positive cells were found in $CREB^{-/-}$ S embryos (not shown). Together, these studies reveal that CREB mutants have defects in the medial and lateral edges as well as in the myotome proper.

The differential severity of myogenic defects in $CREB^{-/-}$ M and $CREB^{-/-}$ S embryos suggests the possibility of differential compensation by ATF-1 and/or CREM. To test this, we examined ATF-1 and CREM expression in wild-type, $CREB^{+/-}$, $CREB^{-/-}$ M and $CREB^{-/-}$ S mice by northern blot analysis (Fig. 2a'). Compared with wild-type embryos, both $CREB^{+/-}$ and $CREB^{-/-}$ S embryos expressed twofold higher levels of CREM, whereas $CREB^{-/-}$ M embryos expressed ~2.5-fold higher levels of ATF-1 and about fourfold higher levels of CREM. We propose that upregulation of

ATF-1 and CREM contributes to the milder myogenic defects observed in $CREB^{-/-}$ M embryos.

We therefore asked whether inhibition of all CREB family members in the somite would block myogenesis *in vivo*. To do this, we generated an adenovirus (Ad) harbouring an expression cassette for a dominant-negative inhibitor of CREB, acidic-CREB (ACREB). ACREB contains an acidic amphipathic extension fused to the amino terminus of the CREB leucine zipper¹⁶ and acts as a dominant-negative for all CREB family members by interfering with their binding to the CRE¹³. We employed microinjection to deliver control and ACREB-expressing adenoviruses, both of which carry a green fluorescent protein (GFP) expression cassette (Ad-GFP and Ad-ACREB, respectively), into E9.75 somites between the forelimb and hindlimb on the right side of the embryo, leaving the left side as an uninjected internal control (Fig. 3a, see Supplementary Video S4). Embryos were then cultured overnight in a rotating bottle culturing unit.

Healthy embryos with robust adenoviral infection in the somite (assessed by GFP, Fig. 3b) were selected for WISH analysis. Control Ad-GFP injected into the somites did not affect *Pax3* ($n = 6$), *MyoD* ($n = 7$) or *Myf5* ($n = 6$) expression (Fig. 3c, e, g). In contrast, myogenesis was inhibited in embryos injected with Ad-ACREB (Fig. 3d, f, h) as indicated by the severe reduction of *Pax3* ($n = 11$), *MyoD* ($n = 10$) and *Myf5* ($n = 13$) expression. Somites anterior and posterior to the injected region showed normal expression of all three genes. The myogenic defects in Ad-ACREB infected somites were more severe than those in the CREB mutant, supporting a redundant function among CREB family members. As *Pax3* and *Myf5* are already expressed at E9.75 (refs 17, 18), we suggest that CREB participates in their maintenance. *Pax1* expression was not altered by ACREB (see Supplementary Fig. S5). Together, these data indicate that CREB-related function is required selectively for myogenic gene expression *in vivo*.

We next addressed how CREB fits into the existing model of myogenic induction. In addition to *Wnt1* and *Wnt7a*, *Shh* has also been implicated in promoting *Myf5* expression^{1,19}. However, because CREB mutants do not have defects in *Pax1*, whose expression depends on *Shh* signalling, it seems unlikely that CREB mediates *Shh* signalling. We therefore tested whether CREB mediates *Wnt1*- and *Wnt7a*-induced myogenesis by taking advantage of an *in vitro* system in which myogenic induction by *Wnt* can be directly assayed. Explants of E9.5 mouse presomitic mesoderm (psm) were co-cultured with control RatB1a cells (B1a), or B1a cells stably expressing *Wnt1* (*Wnt1*-B1a) or *Wnt7a* (*Wnt7a*-B1a) in collagen gels². Expression of *Pax3*, *MyoD* and *Myf5* was assayed 48 h after culture by radioactive *in situ* hybridization (RISH) (Fig. 4a) or semiquantitative polymerase chain reaction with reverse transcription (RT-PCR, Fig. 4b). Although both *Wnt* proteins activated all three myogenic genes, *Wnt1* preferentially activated *Myf5* and *Wnt7a* preferentially activated *MyoD*, consistent with previous reports¹.

Infection of explants with Ad-ACREB repressed *Wnt1*- and *Wnt7a*-mediated myogenesis (Fig. 4c). At low doses, downregulation of *MyoD* was observed. At higher doses, *Myf5* expression was repressed and *Pax3* expression was reduced. Infection with the control Ad-GFP had no effect. These results demonstrate that CREB-related activity is required for *Wnt*-induced myogenesis *in vitro*. Furthermore, different thresholds of CREB activity differentially affect *Pax3*, *MyoD* and *Myf5* expression. ACREB did not affect targets of the canonical WNT/ β -catenin pathway (*Tcf1* and *mTroy*²⁰), indicating the specificity of myogenic inhibition.

To address whether CREB acts downstream of *Wnt*, we used immunohistochemistry to identify whether CREB is phosphorylated in psm co-cultured with *Wnt1*-B1a and *Wnt7a*-B1a cells. We observed moderate levels of P-CREB staining in psm cells from *Wnt1*-B1a co-cultures ($n = 5$, Fig. 4f) and high levels of P-CREB in cells from *Wnt7a*-B1a co-cultures ($n = 5$, Fig. 4g). In contrast, only

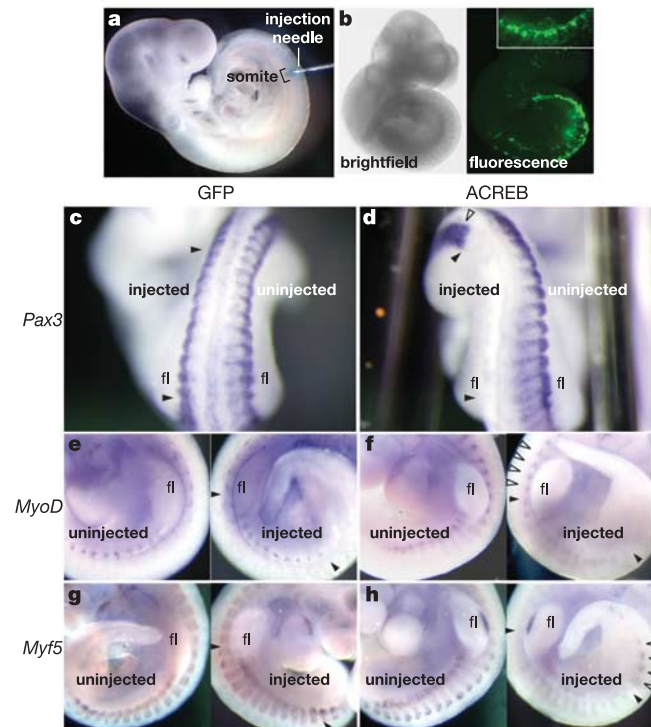


Figure 3 Functional elimination of the CREB family inhibits myogenesis *in vivo*. **a**, GFP-tagged adenoviruses were microinjected into trunk somites at E9.75. **b**, GFP expression (enlarged in inset) in somites following culture (right); brightfield image of same embryo (left). WISH analysis of *Pax3* (**c**, **d**), *MyoD* (**e**, **f**) and *Myf5* (**g**, **h**) in embryos injected with Ad-GFP (GFP) (**c**, **e**, **g**) or Ad-ACREB (ACREB) (**d**, **f**, **h**). Somites on the right side were injected (region of injection located between arrowheads). Uninjected sides and somites anterior and posterior to the injected regions (open arrowheads) retain normal gene expression.

a few scattered P-CREB-positive cells were present in psm alone ($n = 5$, Fig. 4d) and psm/B1a co-cultures ($n = 5$, Fig. 4e). These data suggest that Wnt1 and Wnt7a can activate CREB phosphorylation in psm cells *in vitro*.

We then used cultured cell lines to test if Wnt could induce CREB-mediated transcription. C57MG (known to be Wnt-responsive²¹) and 10T1/2 (known to have myogenic potential²²) cells carrying a CRE-luciferase (CRE-Luc) reporter were co-cultured with B1a, Wnt1-B1a or Wnt7a-B1a cells (Fig. 4h). The reporter activity of C57MG/CRE-Luc cells was induced ~3.7-fold by Wnt1-B1a cells and ~4.3-fold by Wnt7a-B1a cells relative to B1a cells ($n = 3$, $P < 0.005$). Similarly, the reporter activity of 10T1/2/CRE-Luc cells was induced ~2.2-fold by Wnt1-B1a cells and ~3.7-fold by Wnt7a-B1a cells ($n = 3$, $P < 0.005$). Activation of the CRE-Luc reporter by both Wnt proteins appeared to be primarily mediated by the AC cascade, because an inhibitor of AC, 2'-5' dideoxyadenosine (ddA), dramatically reduced reporter activity induced by Wnt1 and Wnt7a (Fig. 4h). Furthermore,

infection of 10T1/2/CRE-Luc cells with Ad-ACREB but not Ad-GFP blocked reporter induction by both Wnt proteins. In contrast, inhibitors of PKC, calcium/calmodulin-dependent protein kinase II or MAP kinase (also known to converge on CREB) had mild or no effect. Wnt-induction of CRE-Luc activity was also independent of other Wnt pathways involving activation of β -catenin or JNK (see Supplementary Fig. S6). Furthermore, CRE-Luc reporter activity was not induced by Shh in either of the cell lines (not shown). Finally, a Wnt11-B1a cell line, which cannot activate *MyoD* and *Myf5* in the psm but can activate the Rho pathway (not shown), failed to induce reporter activity in 10T1/2/CRE-Luc cells ($n = 3$, $P > 0.5$, Fig. 4h). Together, these results demonstrate that Wnt1 and Wnt7a activate CREB-mediated transcription via the AC pathway.

To address whether Wnt activates CREB directly, we treated psm cells with recombinant Wnt3a (rWnt3a) and assayed for P-CREB. We observed CREB phosphorylation within 10 min of application. This induction was dependent on AC and cAMP-dependent PKA

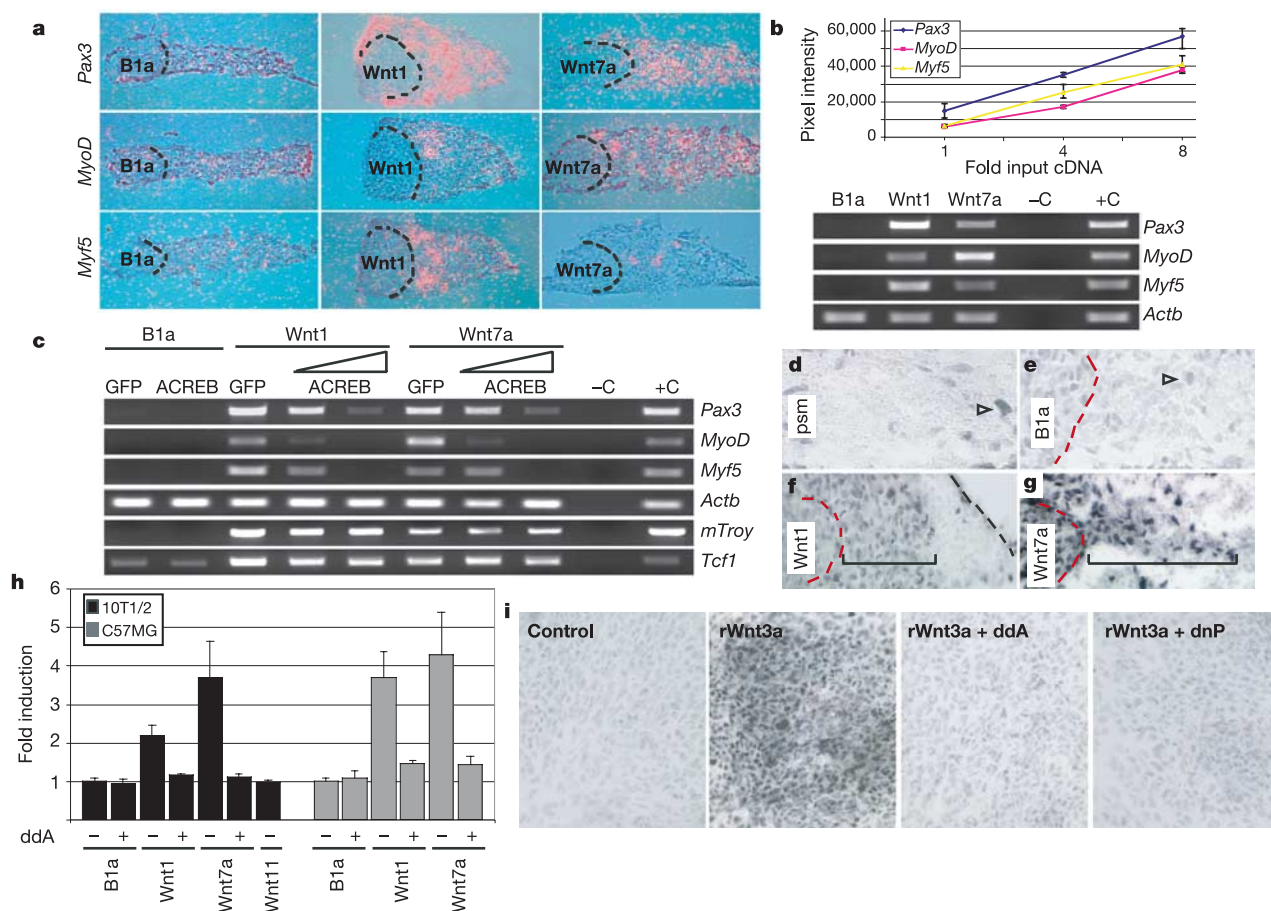


Figure 4 CREB acts downstream of Wnt signalling. **a**, RISH analysis of *Pax3*, *MyoD* and *Myf5* in sections of psm co-cultured with B1a, Wnt1-B1a (Wnt1), or Wnt7a-B1a (Wnt7a) cells. Dashed lines represent the boundary between psm (right) and cell lines (left), pink granules show the RISH signal. **b**, Lower panel shows RT-PCR of *Pax3*, *MyoD* and *Myf5* in co-cultures, cDNA from E10.5 embryos (+C, positive control), no cDNA (-C, negative control) and *Actb* (normalization control). Pixel intensities of typical PCR products are within the linear range of amplification (30 cycles) at noted amounts of cDNA inputs (upper panel). Bars represent range of values ($n = 3$). **c**, RT-PCR of *Pax3*, *Myf5*, *MyoD*, *mTroy* and *Tcf1* in psm co-cultures infected with Ad-GFP (GFP) or increasing amounts of Ad-ACREB (ACREB, relative amounts shown by open triangles). **d–g**, Immunohistochemistry of P-CREB in psm explants. Psm alone (**d**) and B1a cell co-cultures (**e**) have few P-CREB

positive cells (open arrowheads). Wnt-B1a (**f**) or Wnt7a-B1a (**g**) co-cultures induce moderate and strong P-CREB staining, respectively (square brackets). Red dashed lines outline borders (determined by morphology) between psm (right) and cell lines (left), black dashed line in **f** shows the outer edge of psm. P-CREB was also detected in cell lines. **h**, CRE-Luc reporter activities in co-cultures of 10T1/2/CRE-Luc or C57MG/CRE-Luc cells with B1a, Wnt1-B1a, Wnt7a-B1a or Wnt11-B1a cells, with (+) or without (-) 30 μ M ddA treatment. Luciferase activities shown as fold induction relative to B1a co-cultures (mean $\pm$ s.d.). **i** P-CREB staining of psm cells after 10 min treatments: control (carrier media alone), rWnt3a (100 ng ml⁻¹), rWnt3a plus ddA (25 μ M) or rWnt3a plus Ad-dnP (dnP).

activity, because psm cells lacked P-CREB staining when treated with ddA or when infected with an adenovirus carrying a dominant-negative form of the PKA RI α regulatory subunit (dnP, which cannot bind cAMP²³; adenovirus Ad-dnP) ($n = 7$, Fig. 4i). The rapid kinetics of CREB phosphorylation by rWnt3a strongly suggest that CREB is a direct downstream effector of Wnt signalling via AC and PKA in somitic cells.

On the basis of these findings, we propose that the AC signalling cascade leading to CREB activation is required for Wnt-induced myogenesis. To test this, we assessed the effects of pharmacological reagents and adenoviruses that modulate components of this pathway on psm/Wnt-B1a co-cultures. AC activity is regulated by two classes of G proteins: $G_{s\alpha}$, an AC stimulator, and $G_{i\alpha}$, an AC inhibitor²⁴. Increasing AC activity using either pertussis toxin (PTX, a $G_{i\alpha}$ inhibitor) or cholera toxin (CTX, a $G_{s\alpha}$ stimulator), enhanced *Myf5* and *MyoD* expression induced by Wnt1 ($n = 5$) but not by Wnt7a ($n = 5$) (Fig. 5a). *Pax3* levels were relatively unaffected. Conversely, decreasing AC activity using Mas7 (which is 30-fold more potent at activating $G_{i\alpha}$ compared to $G_{s\alpha}$, ref. 25) inhibited both Wnt1- and Wnt7a-induced myogenesis. This repression was not observed in co-cultures exposed to an inactive analogue, Mas17 (Fig. 5b). Wnt1 and Wnt7a protein production was unchanged by Mas7 (not shown). These data imply that modulation of AC activity via G α proteins (which probably couple to Frizzled receptors) may be a mechanism by which Wnt proteins signal.

To test the contribution of AC directly, we applied ddA to the psm explant co-cultures. ddA blocked Wnt1- ($n = 3$) and Wnt7a-induced ($n = 3$) myogenic gene expression (Fig. 5c). ddA did not affect Wnt1 and Wnt7a protein production (not shown). In contrast, forskolin (an activator of AC) enhanced Wnt1- (Fig. 5d) but not Wnt7a- (not shown) mediated myogenesis. In psm/Wnt1-B1a co-cultures ($n = 6$), *Myf5* and *MyoD* expression were respectively

~ 2.3 -fold and ~ 5.7 -fold higher in forskolin-treated samples than in untreated samples. *Pax3* expression was unaffected by forskolin treatment. We observed similar results when we increased levels of cAMP (the downstream effector of AC) using either dibutyl-cAMP (dbc), a membrane-permeable analogue ($n = 3$, Fig. 5d), or 3-isobutyl-1-methyl-xanthine (IBMX), an inhibitor of cAMP phosphodiesterase (not shown). In contrast, activation of PKC by 12-*O*-tetradecanoylphorbol-13-acetate (TPA) ($n = 3$, Fig. 5d) had no effect. Together, these data show that although both Wnt proteins require AC activity, they may differ in their capacity to activate AC in the psm (Fig. 4f, g), providing a possible explanation for their preferential myogenic-inducing properties.

We next asked whether blocking PKA activity would block myogenesis. Infection of explant co-cultures with Ad-dnP, but not control Ad-GFP, repressed both Wnt1- and Wnt7a-directed myogenic gene expression (Fig. 5e). Wnt1 and Wnt7a protein production and secretion by Wnt1-B1a and Wnt7a-B1a cells were not affected by dnP, as verified by normal myogenic induction in co-cultures of uninfected psm with Ad-dnP infected Wnt1-B1a or Wnt7a-B1a cells (not shown). We did not detect *Pax1* expression in these cultures, indicating that unlike what has been reported in the neural tube²⁶, the Shh pathway is not activated by dnP in the mouse somitic mesoderm. Finally, consistent with Ad-ACREB injections, embryos injected with Ad-dnP also exhibited loss of *Pax3*, *MyoD* and *Myf5* expression ($n = 12$ each, Fig. 5f–h). These results together indicate that PKA activity is required for Wnt-mediated myogenesis *in vitro* and *in vivo*.

Our data support a model in which Wnt-mediated myogenic gene expression is dependent on the activation of AC, PKA and CREB. Consistent with our model, core CRE sites (CGTCA) are present in the promoter regions of *Myf5* and *Pax3* genes. Whether CREB directly activates the myogenic genes remains to be deter-

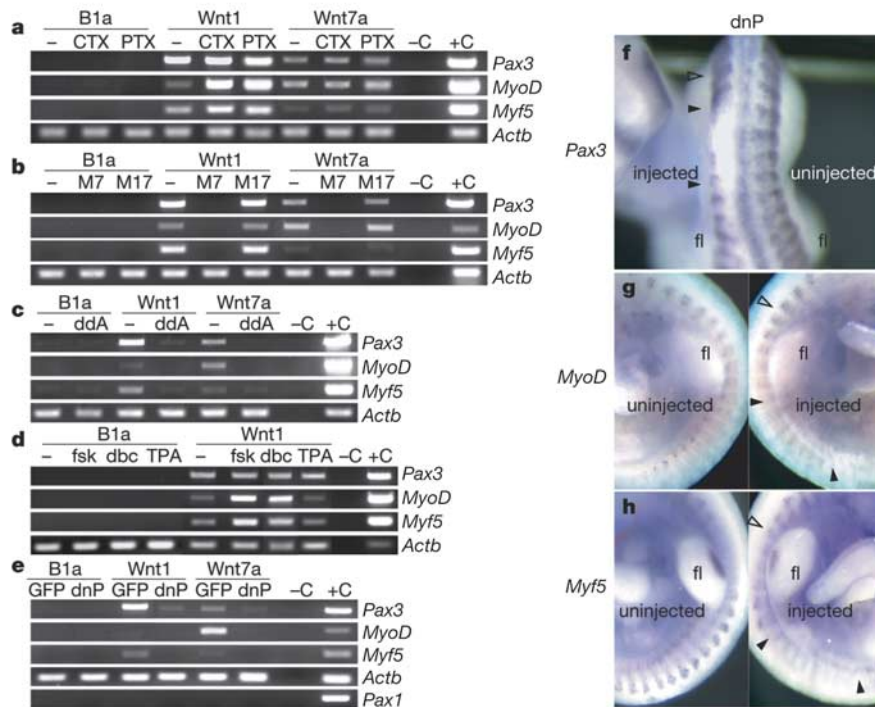


Figure 5 G proteins, AC, cAMP and PKA mediate Wnt-induced myogenesis. **a–e**, RT-PCR analysis of *Pax3*, *MyoD* and *Myf5* in psm co-cultures with B1a, Wnt1-B1a (Wnt1) and Wnt7a-B1a (Wnt7a) cells treated with **(a)** CTX (50 ng ml⁻¹) or PTX (75 ng ml⁻¹), **(b)** Mas7 (25 μ M) or Mas17 (25 μ M), **(c)** ddA (25 μ M), **(d)** forskolin (fsk, 5 μ M), dbc (1 mM) or TPA (100 ng ml⁻¹) and **(e)** Ad-GFP (GFP) or Ad-dnP (dnP); control samples as in Fig. 4,

untreated cells (–). **e**, *Pax1* expression is not activated by dnP. These treatments have no effect on B1a co-cultures. **f–h**, WISH analysis of *Pax3*, *MyoD* and *Myf5* in Ad-dnP injected embryos. Black arrowheads bracket targeted somites. Uninjected sides and somites anterior and posterior to the injected regions (open arrowheads) retain normal gene expression.

mined. The new link between Wnt and PKA/CREB signalling broadens the repertoire of downstream components used by Wnt proteins. As certain Wnt proteins are also known to participate in cell proliferation and survival³, they may utilize the PKA/CREB pathway to mediate these processes because these roles are well described for CREB²⁷. Given the diverse functions of the Wnt family in vertebrate and invertebrate embryogenesis, our data provide the basis for investigating a general role of PKA/CREB in Wnt-directed developmental programs. Our discovery may also encourage the manipulation of PKA/CREB pathway components as a means to modulate or re-direct Wnt-mediated stem cell renewal and cancer. □

Methods

Explant culture

E9.5 CD1 mouse psm explants were cultured in collagen gel alone, or in combination with B1a, Wnt1-B1a or Wnt7a-B1a cells as described², except that 2.5 ng ml⁻¹ basic fibroblast growth factor (bFGF, Promega) was used. All cultures were 48 h. For rWNT3a application, psm explants were plated on 8-chamber slides (Falcon) for 18 h in the presence or absence of pharmacological reagents or adenovirus, followed by 10 min of treatment with rWNT3a (R&D Systems) or carrier media (control). Mas7, Mas17, CTX, PTX, dbc, IBMX, TPA and ddA were obtained from Calbiochem. Forskolin was obtained from Sigma. All reagents were directly applied to cultures.

Adenovirus production

GFP adenovirus (Ad-GFP) was previously described²⁸. FLAG-tagged-ACREB¹⁶ and dnP²³ expressing adenoviruses (Ad-ACREB and Ad-dnP, respectively) were produced using the AdenoEasy System (Qbiogene). Purified Ad-dnP and Ad-ACREB were verified to be functional. 10⁸–10¹⁰ plaque-forming units were added to explant culture media to achieve ~100% infection, as visualized by GFP fluorescence.

Semi-quantitative RT-PCR

Total RNA from the explants was isolated using RNasol and analysed by RT-PCR with 30–32 cycles of amplification. For primer sequences, see (<http://www.ciwemb.edu/labs/fan/index.html>). PCR products were resolved on 2% agarose gels and visualized by ethidium bromide staining. Images were captured using a UVP 7500 Gel Documentation System and quantified using Image-Quant v.1.2 (Amersham Pharmacia). Each marker was assayed in an independent PCR reaction per cDNA sample. Quantified PCR products were normalized to β -actin (*Actb*) intensity of the same sample and then used for quantitative comparison between samples.

Immunohistochemistry

Explants and embryos at E9.5 and E10.5 were fixed in 4% paraformaldehyde in PBS and processed for paraffin sectioning at 14 μ m. Sections were incubated with primary antibodies against CREB (1:500), P-CREB (1:500) (Upstate Biotechnology), Pax3 C-20 (1:500), MyoD (1:500) or Myf5 (1:750) (Santa Cruz Biotechnology). Biotinylated secondary antibodies (1:200) and VectaStain Elite ABC Reagent (Peroxidase Kit, Vector Labs) were then used and staining was visualized using 3,3'-diaminobenzidine substrate (Vector Labs).

Mice

CREB^{+/-} mice were crossed to obtain mutant embryos. Yolk sac DNA was used for PCR genotyping as described¹⁴.

WISH and RISH

WISH was performed using digoxigenin (DIG)-UTP (Roche) labelled antisense RNA probes as described²⁸. Signal was detected using an alkaline-phosphatase-conjugated anti-DIG antibody and NBT/BCIP substrate (Roche). Probes used were: Pax1 (ref. 28), Pax3 (ref. 28), Myf5 (ref. 28) and MyoD (nucleotides 795–1166 of GenBank accession number M84918). RISH was performed on sectioned explants as described² using the above probes labelled with ³⁵S-UTP.

Northern analysis

Total RNA was isolated from CREB^{+/+}, CREB^{+/-}, CREB^{-/-} M and CREB^{-/-} S embryos at E11.5 using TRIzol Reagent (Invitrogen) and quantified using a spectrophotometer. 10 μ g RNA per sample was used for standard northern blot analysis. CREM (from J. Blendy) and ATF-1 probes were labelled with ³²P-CTP using the Prime II Kit (Stratagene). Hybridization signals were quantified using a phosphorimager and ImageQuant software.

Embryo injection and culture

Whole E9.75 CD1 embryos were collected in L-15/5% heat inactivated horse serum (Invitrogen) for injection. Purified adenoviruses were injected into somites (4.6–9.2 nl per somite) using a Nanoinjector (Drummond Scientific). Following injection, embryos were transferred to a rotating bottle culture unit (BTC Engineering) and cultured overnight in media (2 ml per embryo) consisting of 100 units ml⁻¹ penicillin, 100 μ g ml⁻¹ streptomycin, 50% DMEM/F12 (Invitrogen) and 50% freshly isolated, sterile-filtered rat serum, under humidified conditions at 40% O₂, 5% CO₂, 55% N₂, 37 °C.

Reporter cell lines

10T1/2 and C57MG cell lines each carrying a stably integrated CRE-Luc reporter were clonally selected with hygromycin (200 μ g ml⁻¹). For co-cultures, 10T1/2 or C57MG cells were mixed (at 1:2 and 2:1 ratios, respectively) with B1a or Wnt-B1a cells. Luciferase activity was determined as described²⁸. The protein concentration used for normalization was determined using the BCA protein assay reagent (Pierce).

Received 5 August; accepted 20 October 2004; doi:10.1038/nature03126.

Published online 28 November 2004.

1. Tajbakhsh, S. *et al.* Differential activation of *Myf5* and *MyoD* by different Wnts in explants of mouse paraxial mesoderm and the later activation of myogenesis in the absence of *Myf5*. *Development* **125**, 4155–4162 (1998).
2. Fan, C. M., Lee, C. & Tessier-Lavigne, M. A role for WNT proteins in induction of dermomyotome. *Dev. Biol.* **191**, 160–165 (1997).
3. Moon, R., Bowerman, B., Boutros, M. & Perrimon, N. The promise and perils of Wnt signaling through β -catenin. *Science* **296**, 1644–1646 (2002).
4. Ordahl, C. (ed.) *Somitogenesis* Part 2 129–268 (Academic, Toronto, 2000).
5. Rudnicki, M. *et al.* *MyoD* or *Myf5* is required for the formation of skeletal muscle. *Cell* **75**, 1351–1359 (1993).
6. Tajbakhsh, S., Rocancourt, D., Cossu, G. & Buckingham, M. Redefining the genetic hierarchies controlling skeletal myogenesis: *Pax3* and *Myf5* act upstream of *MyoD*. *Cell* **89**, 127–138 (1997).
7. Parr, B., Shea, M., Vassileva, G. & McMahon, A. Mouse Wnt genes exhibit discrete domains of expression in the early embryonic CNS and limb buds. *Development* **119**, 247–261 (1993).
8. Capdevila, J., Tabin, C. & Johnson, R. Control of dorsoventral somite patterning by Wnt-1 and β -catenin. *Dev. Biol.* **193**, 182–194 (1998).
9. Barnes, M., Duckworth, D. & Beeley, L. Frizzled proteins constitute a novel family of G protein-coupled receptors, most closely related to the secretin family. *Trends Pharmacol. Sci.* **19**, 399–400 (1998).
10. Kuhl, M., Sheldahl, L., Park, M., Miller, J. & Moon, R. The Wnt/Ca²⁺ pathway: a new vertebrate Wnt signaling pathway takes shape. *Trends Genet.* **16**, 279–283 (2000).
11. Huelsken, J. & Behrens, J. The Wnt signalling pathway. *J. Cell Sci.* **115**, 3977–3978 (2002).
12. Wang, H. Y. & Malbon, C. Wnt-frizzled signaling to G-protein-coupled effectors. *Cell. Mol. Life Sci.* **61**, 69–75 (2004).
13. Shaywitz, A. & Greenberg, M. CREB: a stimulus-induced transcription factor activated by a diverse array of extracellular signals. *Annu. Rev. Biochem.* **68**, 821–861 (1999).
14. Rudolph, D. *et al.* Impaired fetal T cell development and perinatal lethality in mice lacking the cAMP response element binding protein. *Proc. Natl Acad. Sci. USA* **95**, 4481–4486 (1998).
15. Ikeya, M. & Takada, S. Wnt signaling from the dorsal neural tube is required for the formation of the medial dermomyotome. *Development* **125**, 4969–4976 (1998).
16. Ahn, S. *et al.* A dominant-negative inhibitor of CREB reveals that it is a general mediator of stimulus-dependent transcription of c-fos. *Mol. Cell. Biol.* **18**, 967–977 (1998).
17. Ott, M., Bober, E., Lyons, G., Arnold, H. & Buckingham, M. Early expression of the myogenic regulatory gene, *Myf5*, in precursor cells of skeletal muscle in the mouse embryo. *Development* **111**, 1097–1107 (1991).
18. Bober, E., Franz, T., Arnold, H., Gruss, P. & Tremblay, P. *Pax3* is required for the development of limb muscles: a possible role for the migration of dermomyotomal muscle progenitor cells. *Development* **120**, 603–612 (1994).
19. Munsterberg, A., Kitajewski, J., Bumcrot, D., McMahon, A. & Lassar, A. Combinatorial signaling by Sonic hedgehog and Wnt family members induces myogenic bHLH gene expression in the somite. *Genes Dev.* **9**, 2911–2922 (1995).
20. Buttitta, L., Tanaka, T., Chen, A., Ko, M. & Fan, C. M. Microarray analysis of somitogenesis reveals novel targets of different WNT signaling pathways in the somitic mesoderm. *Dev. Biol.* **258**, 91–104 (2003).
21. Mason, J., Kitajewski, J. & Varmus, H. Mutational analysis of mouse Wnt-1 identifies two temperature-sensitive alleles and attributes of Wnt-1 protein essential for transformation of a mammary cell line. *Mol. Biol. Cell* **3**, 521–533 (1992).
22. Weintraub, H. *et al.* Activation of muscle-specific genes in pigment, nerve, fat, liver, and fibroblast cell lines by forced expression of *MyoD*. *Proc. Natl Acad. Sci. USA* **86**, 5434–5438 (1989).
23. Clegg, C., Correll, L., Cadd, G. & McKnight, G. S. Inhibition of intracellular cAMP-dependent protein kinase using mutant genes of the regulatory type I subunit. *J. Biol. Chem.* **262**, 13111–13119 (1987).
24. Manning, D. G. *Proteins: Techniques of Analysis* (Florida, CRC, 1999).
25. Iyengar, R. *Methods in Enzymology: Heterotrimeric G Proteins* Vol. 237 26–37 (Academic, San Diego, 1994).
26. Epstein, D., Marti, E., Scott, M. & McMahon, A. Antagonizing cAMP-dependent protein kinase A in the dorsal CNS activates a conserved Sonic hedgehog signaling pathway. *Development* **122**, 2885–2894 (1996).
27. Lonze, B. & Ginty, D. Function and regulation of CREB family transcription factors in the nervous system. *Neuron* **35**, 605–623 (2002).
28. Buttitta, L., Mo, R., Hui, C. C. & Fan, C. M. Interplays of *Gli2* and *Gli3* and their requirement in mediating Shh-dependent sclerotome induction. *Development* **130**, 6233–6243 (2003).

Supplementary Information accompanies the paper on www.nature.com/nature.

Acknowledgements We thank D. Koshland, Y. Zheng, A. Fire, L. Buttitta and M. Tessier-Lavigne for critical reading of this manuscript and C. Pratt for graphics assistance. D.D.G. is an investigator of the Howard Hughes Medical Institute.

Competing interests statement The authors declare that they have no competing financial interests.

Correspondence and requests for materials should be addressed to C.M.F. (fan@ciwemb.edu).

An autoregulatory circuit for long-range self-organization in *Dictyostelium* cell populations

Satoshi Sawai, Peter A. Thomason & Edward C. Cox

Department of Molecular Biology, Princeton University, Princeton, New Jersey 08544, USA

Nutrient-deprived *Dictyostelium* amoebae aggregate to form a multicellular structure by chemotaxis, moving towards propagating waves of cyclic AMP that are relayed from cell to cell. Organizing centres are not formed by founder cells, but are dynamic entities consisting of cores of outwardly rotating spiral waves^{1–4} that self-organize in a homogeneous cell population. Spiral waves are ubiquitously observed in chemical reactions as well as in biological systems^{5–8}. Although feedback control of spiral waves in spatially extended chemical reactions has been demonstrated in recent years^{9,10}, the mechanism by which control is achieved in living systems is unknown. Here we show that mutants of the cyclic AMP/protein kinase A pathway show periodic signalling, but fail to organize coherent long-range wave territories, owing to the appearance of numerous spiral cores. A theoretical model suggests that autoregulation of cell excitability mediated by protein kinase A acts to optimize the number of signalling centres.

We have studied the spatio-temporal geometry of chemoattractant waves in *Dictyostelium* mutants related to the protein kinase A (PKA) pathway (Fig. 1a). *Dictyostelium* cells are excitable, and when stimulated with extracellular cyclic AMP (cAMP), they activate adenylyl cyclase (ACA) thus producing more cAMP, which in turn stimulates neighbouring cells (see ref. 11 for details). This excitability is mainly determined by two factors. One is the sensitivity of the cell to cAMP, which increases as expression of the cAMP receptor CAR1 rises after starvation. The other is the level of cAMP output, which depends on ACA activity. It is known that CAR1, ACA and other early components, some of which mediate the coupling between CAR1 and ACA, are induced by pulses of extracellular cAMP acting through intracellular cAMP and protein kinase A (PKA)¹² (Fig. 1a).

Signalling between *Dictyostelium* cells on an agar substrate is easily visualized by means of dark-field optics^{13,14}, in which propagating waves of cAMP appear as optical density waves. Wild-type amoebae initially propagate waves that have spiral and circular geometries, and by 5 h, spiral wave territories (diameter ~4 mm) encompassing a few wavelengths dominate and persist for 1–2 h (Fig. 1b; WT). Streams of aggregating amoebae begin to form as the signal diminishes. In marked contrast, cells mutant in the mitogen-activated protein kinase gene, *erkB*, and in the internal cAMP phosphodiesterase gene, *regA*, do not propagate the trains of waves observed in wild-type cells (Fig. 1b; *erkB*, *regA*). Surprisingly, when examined on sub-millimetre scales, strong periodic oscillations are evident everywhere (Fig. 1c), and closer examination reveals that the mutants are capable of propagating waves over short distances of approximately one wavelength (~1 mm).

To quantify the oscillations, we performed wavelet analysis¹⁵. Oscillatory signalling in wild-type cells begins ~3–4 h after plating, has a periodicity of 5–7 min, and continues for ~3 h (Fig. 2; AX2, DH1). Periodic signalling by *erkB* mutants begins late, at ~6 h after starvation, starting out with a 5–7-min periodicity, then slowing down to a 10–11-min periodicity. They continue signalling for more than 8 h, without aggregating, as evidenced by the extended wavelet portrait (Fig. 2; *erkB*). *RegA* mutants oscillate with an ~7-min periodicity beginning 2–3 h after starvation and lasting for ~2 h

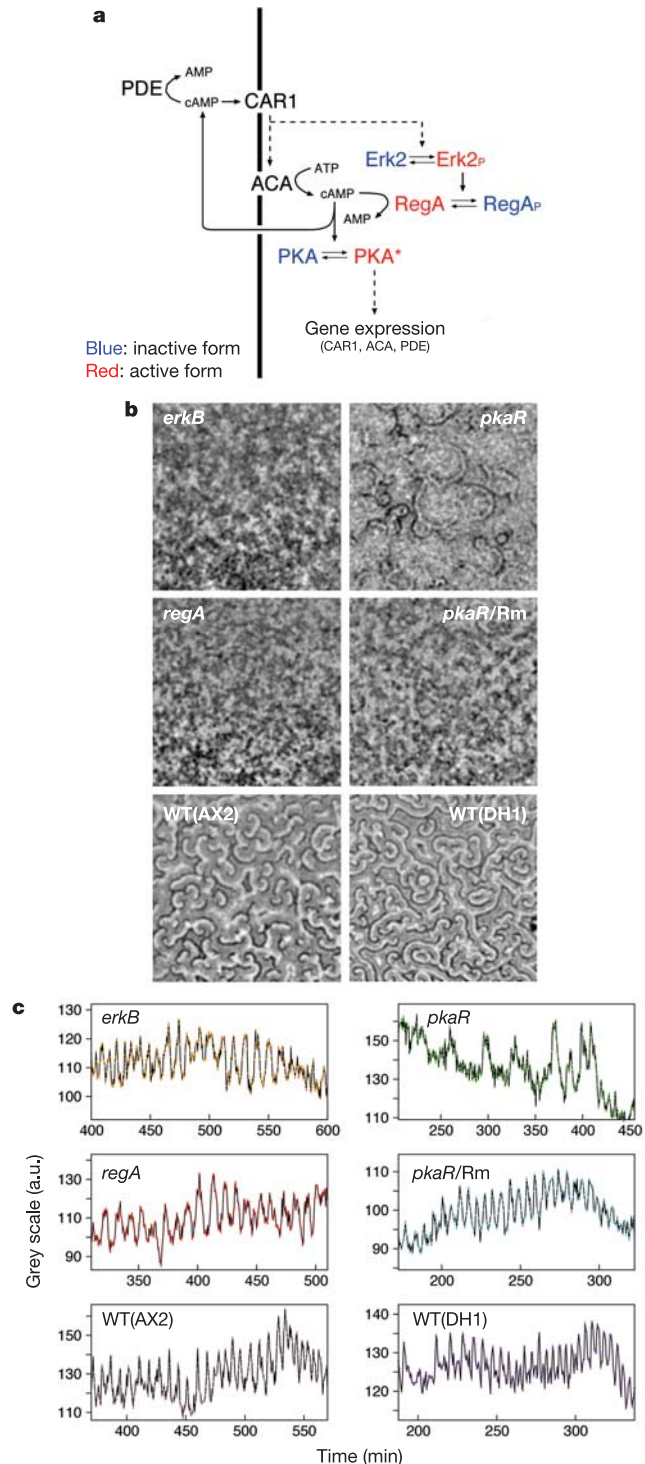


Figure 1 Mutants in the PKA circuit show spatially aberrant wave patterns. **a**, Upon cAMP stimulation, newly synthesized cAMP is secreted and stimulates other cells. It is then degraded by extracellular phosphodiesterase (PDE)²². The mitogen-activated protein kinase Erk2 (ref. 20) is transiently phosphorylated upon cAMP stimulation and regulates cAMP degradation by inhibiting the intracellular phosphodiesterase RegA²¹. PKA mediates pulse-induced transcription of genes encoding CAR1, PDE and ACA¹². Asterisk PKA is the dissociated form of the catalytic subunit. **b**, Optical density waves. *erkB*, *regA* and *pkaR* are knockout mutants of Erk2, RegA and the regulatory subunit of PKA, respectively; *pkaR/Rm* is a *pkaR* mutant expressing a dominant inhibitor form of the subunit. Frame-subtracted images are from 29 × 29 mm areas taken 4–6 h after starvation; images of *erkB* were taken at 10 h. See Supplementary Video S1. **c**, Mutants and wild-type cells display ~7-min period optical density oscillations (representative time-plots from arbitrary 0.18 × 0.18 mm regions). Grey scale, arbitrary units (a.u.).

(Fig. 2; *regA*), after which they aggregate to form minute mounds and morphologically aberrant fruiting bodies, as previously described¹⁶. These developmental anomalies, as well as other features discussed below, were corrected in complemented strains (*erkB*+/+, *regA*+/+). Although the developmental timing of the oscillations therefore varies, the kinetics of the oscillations are very similar in all strains.

The core of a spiral wave in excitable media is a singular point in space that neighbours every phase in the oscillation cycle. By using a time delay embedding technique⁸, the cores could be rigorously located (Fig. 3a; see Supplementary Information). As shown in Fig. 3b, *erkB* and *regA* mutants display ~twofold and ~fourfold increases in the number of singularities, respectively. These cores, with the exception of a few per cent that become entrained by an adjacent spiral, persist for more than 1 h in both mutants and the wild type. Thus the inability of *regA* and *erkB* mutants to propagate signals over long distances is caused by an excess of spiral cores, not by a failure to establish them; waves emanating from many cores collide and annihilate each other before they can propagate very far.

cAMP spirals in *Dictyostelium* take the form of an involute of a circle³. Hence, there is an upper limit on the number of cores per unit area, resulting from a geometrical constraint on spirals in excitable media. The smallest circle, which encloses a counter-rotating spiral pair of one wavelength, occupies roughly an area $\pi(2\pi r + 2r)^2$, where r is the core radius measured from the motion of singularities ($r \sim 0.1$ mm). At the closest packing of these circles, a maximum of ~110 cores per cm^2 is allowed, which is close to what is observed in *regA* mutants (Fig. 3b). These maximally packed phase singularities help to explain an earlier observation that wild-type cells placed in a territory of *regA* mutant cells appeared to respond to multiple signalling centres, which were present even within a small area¹⁷.

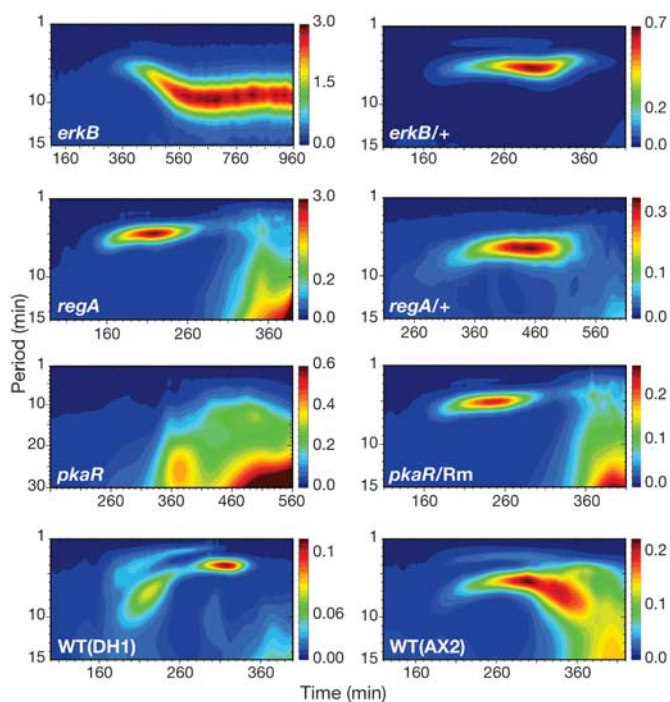


Figure 2 Time evolution of signalling is greatly altered in the mutants. The periodicity of the signal at time t (min) after starvation is represented by the scale s (min) that gives the peak wavelet amplitude $W(s, t)$ (contour plot colour axis; see Supplementary Methods). WT cells reach maximum wavelet amplitude at 4.5 h after starvation, and signalling lasts 3 h; *erkB* reaches its peak activity 9.5 h after starvation and continues to signal for more than 10 h; *regA* takes 3 h to reach the maximum amplitude, and signalling lasts for 1.5 h. All experiments were repeated at least three times; representative results are shown.

Is wave turbulence in *erkB* and *regA* populations associated with the altered activity of PKA (Fig. 1a)? A null mutant of the regulatory (R) subunit of PKA (*pkaR*; having unrestrained PKA activity) shows a few initial spikes, each of which takes ~20–30 min to recover to the basal level (Fig. 1c). These spikes initiate circular waves that propagate in space with a plateau-like form (having the appearance of raindrops on a surface, after frame subtraction; Fig. 1b) because the cells return very slowly to their original brightness after excitation. After a few of these spikes, there is a transition to almost wild-type ~10-min oscillations, which last for 1–2 h (Fig. 1c; *pkaR* from ~400 min onwards), during which time cells aggregate to form small mounds. These later waves appear to be circular, but they are difficult to analyse because the cells aggregate precociously. This complication was also observed for cells expressing the catalytic subunit of PKA under the *act15* promoter (data not shown), as would be expected from its constitutive PKA activity. The overall

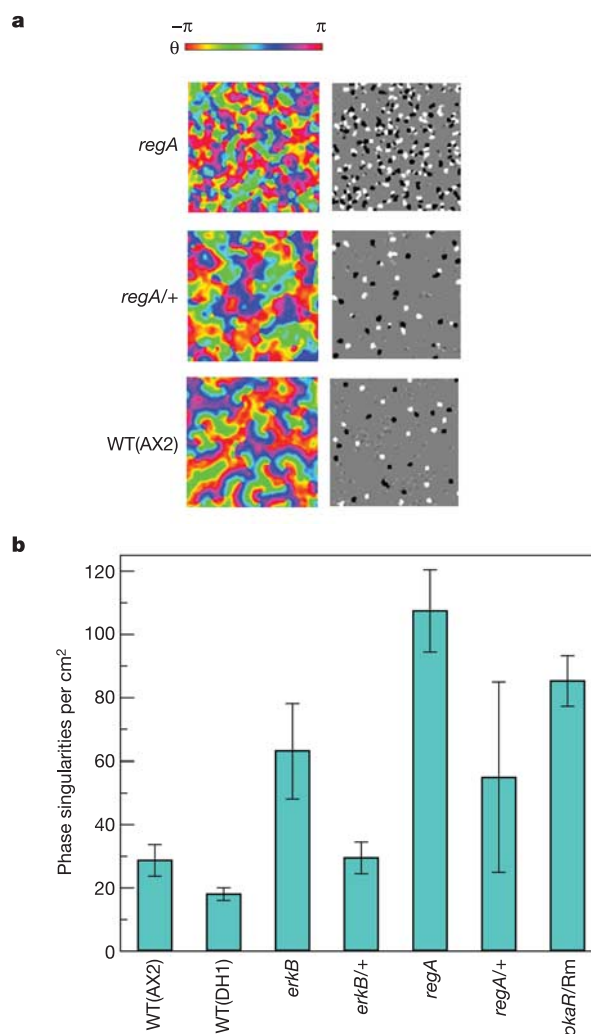


Figure 3 Time delay embedding and the extraction of spatial phase singularities⁸ reveals the unusual number of spiral cores in the mutants (see Supplementary Methods, Video S2 and Fig. S4). **a**, Left panels: the angular variable representation of the optical density waves. *regA*+/+, complemented mutant strain. Spiral cores appear as singularities where all colours collapse into a single point. Right panels: time-average images of over 50 min of phase singularities, which equal either 0, $+2\pi$, or -2π , represented as grey, black and white, respectively. $+2\pi$, a clockwise rotating spiral; -2π , an anti-clockwise rotating spiral. Panels represent 12×12 mm regions. **b**, Number of phase singularities per cm^2 in mutants and wild type. Error bars represent standard deviations for $n = 3-4$ identical experiments.

wavelet pattern of *pkaR* cells is shifted towards longer periodicity compared with the parental strain (Fig. 2; *pkaR* and DH1).

We next introduced a dominant inhibitor form (Rm) of the PKA regulatory subunit¹⁸ into the *pkaR* mutant. This regulatory subunit has severely impaired cAMP binding, and thus remains associated with the PKA catalytic subunit (PKA-C), inhibiting it even in the presence of cAMP¹⁸. Figure 1b, c shows that a *pkaR* mutant expressing Rm (see Methods) exhibits waveforms that are spatially turbulent but that oscillate periodically in time. The periodicity of the oscillations is 5–7 min, and the signal lasts for 2 h (Fig. 2). The number of phase singularities is ~fourfold higher than the parental wild-type DH1 cells (Fig. 3b and Supplementary Fig. S4b). As a control, a further modified form of PKA-R, whose site of interaction with PKA-C is also deleted (Rc)¹⁸ was used. A *pkaR* mutant expressing Rc showed no change from the parental *pkaR* mutant in the wave phenotype (data not shown).

These observations indicate that the PKA pathway (Fig. 1a) is dispensable for optical density oscillations, but is necessary for the establishment of long-range coherent wave territories. The requirement of ACA for the optical density oscillations has been conclusively demonstrated using a temperature-sensitive mutant of ACA¹⁹. The presence of sustained oscillations in *erkB* populations is therefore surprising, given that they show a strongly reduced cAMP relay response²⁰ in a biochemical assay. This may be due to excess cAMP degradation by uninhibited RegA in the assay. Erk2 is not necessary for activation of ACA, as demonstrated with the double mutant *regA/erkB*²¹. Moreover, an application of a cAMP mist to the *erkB* mutants resets waves and causes bulk oscillations (see Supplementary Fig. S5), as expected from the phase-resetting property of cAMP oscillations. Nor is there periodic signalling in a null mutant of the extracellular cAMP phosphodiesterase²² (data not

shown). Together with the finding that *regA* and *pkaR* mutants are able to oscillate, these facts suggest that periodic cAMP signalling is mainly achieved not by an intracellular feedback loop through PKA²³, but by an extracellular one²⁴ in which cAMP is degraded by the extracellular phosphodiesterase (see Supplementary Discussion).

How could a defective PKA pathway result in the waveforms we have observed? We studied this problem using a cellular automaton model²⁵ that incorporates a cAMP pulse-dependent and pulse-independent increase in excitability (see equations in Supplementary Information). Our simulations show that when the cAMP-induced increase in cell excitability falls outside a certain range—where the feedback strength β is too high or too low—excess phase singularities appear (Fig. 4a). On the basis of the mutant wave phenotype and this generic qualitative feature of the model, we propose that wave territory size is optimized by pulse-induced evolution of excitability through RegA, Erk2 and PKA activity (Fig. 1a). When feedback is entirely absent (that is $\beta = 0$, as with the *pkaR* mutant), and the excitability is quickly elevated in a cell-autonomous manner, spiral waves fail to form, and circular waves are observed instead (Fig. 4b). In *erkB* mutants, the progression of excitability is weakly coupled to cAMP stimuli (Fig. 4a; $\beta < 0.002$), owing to low cAMP concentration. Similarly, *pkaR*/Rm mutants may also be working in this regime. *RegA* mutants, however, fail to bring the cAMP concentration down sufficiently after each stimulus, so that excitability increases too rapidly (Fig. 4a; $\beta > 0.02$).

The model predicts that the mechanism by which excitability evolves through time is important in determining waveform. We examined this by first starving cells in suspension for various time periods before plating them for wave analysis. Cells starved for 4–8 h generated spiral waves, although they initiated waves within 2 h of

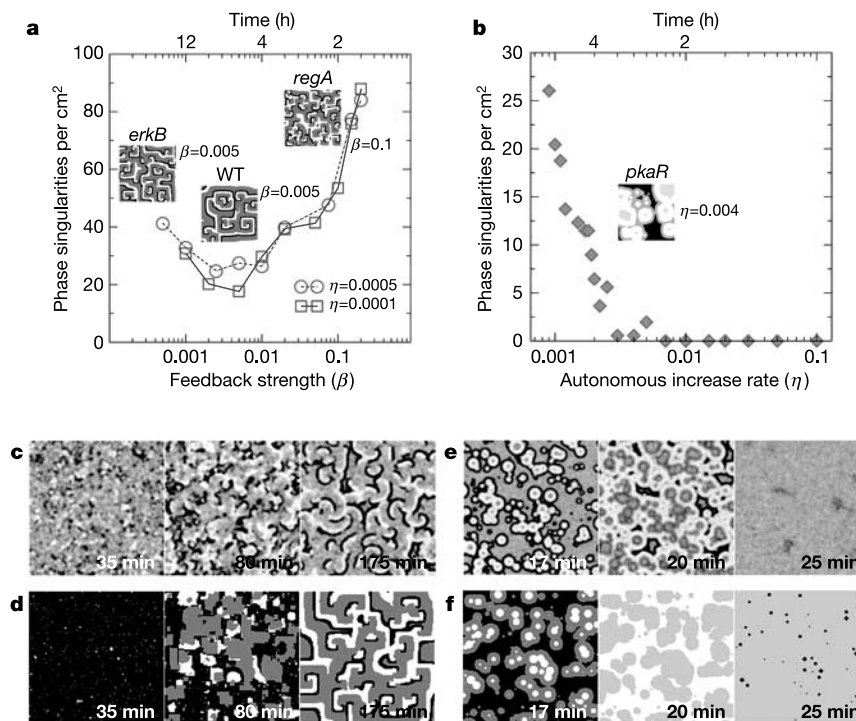


Figure 4 Cellular automaton simulations and resulting waveforms. **a**, When the autonomous increase in excitability (η) is suppressed ($\eta < 0.001$), the pulse-induced increase in excitability (feedback strength β) holds back the emergence of singularities. Insets are representative snapshots of asymptotic states. β and η were chosen to fit the pattern-formation timing observed in the experiments (indicated on the upper x axis). **b**, When feedback is absent ($\beta = 0$), high autonomous increase in excitability

($\eta > 0.002$) precludes formation of spiral waves. Wild-type AX2 cells were starved for 6.5 h (**c**) and 8 h (**e**) before observation submerged in PB. Simulations from an intermediate value of excitability (E) at time zero, $E(0) = 0.5$ (**d**) and near maximum $E(0) = 0.8$ (**f**): $\beta = 0.01$, $\eta = 0.0005$; other parameters are as in **a** (see Supplementary Equations). Consistency with experimental data (**c**, **e**; see Supplementary Video S3) demonstrates that the parameters in **a** are well constrained.

plating (Fig. 4c). Cells starved for ~8 h or longer showed bulk oscillations immediately after plating (Fig. 4e). They failed to generate spatial phase singularities even after signalling for 6 h. This is in good agreement both with our simulations (Fig. 4d, f; see also ref. 25) and with an earlier finding that cells that are more than 8 h into development are unable to recover spiral waves after being reset by applying a mist of cAMP²⁶.

This is the first experimental report on a genetic circuit that fine-tunes spontaneous symmetry-breaking of a self-organizing nature. Although many of the molecular details of the network remain to be elucidated, this minimal model captures the essence of our current understanding of the feedback and successfully describes the observed spatial features of waves propagating in wild-type and mutant cells. Similar autoregulatory machineries may be at work in other excitable systems, such as heart muscle⁸, to prevent deleterious occurrence of spiral waves. □

Methods

Cell culture and dark-field wave observation

Wild-type and mutant strains of *Dictyostelium discoideum* were grown with *Escherichia coli* B/r on glucose–yeast–peptone plates²⁷ or grown axenically in HL5 medium at 22 °C. Vegetative cells (4.5×10^7) were washed with 20 mM KH₂PO₄/Na₂HPO₄ buffer (PB, pH 6.5) and plated on 1% PB–agar (Bacto-agar, Gibco) plates (10 cm diameter) unless otherwise noted. Cells were allowed to settle for 15 min before the supernatant was removed, and plates allowed to dry for 15 min in a sterile hood. Dark-field optics recordings were performed at 22 °C as previously described¹. Digitized images were taken every 30 s using an image grabber (Scion LG-3). The images are 8-bit grey scale, 640 × 480 pixels with a resolution of 61 µm per pixel. They were integrated at 30 Hz over 0.3 s for noise reduction.

Mutant strains

ErkB (HS174; ref. 20) and *regA* (HM1015; ref. 16) mutants are insertional mutants generated by homologous recombination. *RegA*+/+ (HM2041) is HM1015 transformed with a vector expressing *regA* driven by its own promoter. *ErkB*+/+ (HS174-A1) is HS174 expressing *erkB* under the *act15* promoter. All mutant and rescued strains were verified during use. The *pkaR* mutant (MYC1002; ref. 16) is an insertional mutant obtained by restriction enzyme mediated integration. We also tested a classical *pkaR* mutant generated by chemical mutagenesis (HTY506; ref. 28) and confirmed that overall wave phenotype was indistinguishable from MYC1002 (data not shown). For expression of PKA-Rm in MYC1002, transformants were isolated in HL5 medium (8 µg ml⁻¹ G418) and then on bacterial plates (*E. coli* B/r-1 on glucose–yeast–peptone containing 50 mg ml⁻¹ G418). Expression of PKA-R in *pkaR*/Rm clones was confirmed by western blots probed with a monoclonal antibody raised against *Dictyostelium* PKA-R²⁸. Two independent clonal isolates of *pkaR*/Rm that showed expression levels comparable to that of wild-type DH1 developed for 4–8 h were used for the wave analysis. Three other isolated clones expressed higher levels of the subunit, showed no sign of periodic oscillations and could not aggregate. We also tested a null mutant of PKA-C (*pkaC*) and observed the same phenotype (data not shown). DH1 and AX2 are parental wild-type strains for MYC1002 and HM1015, respectively.

Received 21 June; accepted 23 November 2004; doi:10.1038/nature03228.

1. Palsson, E. *et al.* Selection for spiral waves in the social amoebae *Dictyostelium*. *Proc. Natl Acad. Sci. USA* **94**, 13719–13723 (1997).
2. Lauzeral, J., Halloy, J. & Goldbeter, A. Desynchronization of cells on the developmental path triggers the formation of spiral waves of cAMP during *Dictyostelium* aggregation. *Proc. Natl Acad. Sci. USA* **94**, 9153–9158 (1997).
3. Foerster, P., Muller, S. & Hess, B. Curvature and spiral geometry in aggregation patterns of *Dictyostelium discoideum*. *Development* **109**, 11–16 (1990).
4. Siegfert, F. & Weijer, C. Digital image processing of optical density wave propagation in *Dictyostelium* and analysis of the effects of caffeine and ammonia. *J. Cell Sci.* **93**, 325–335 (1989).
5. Lechleiter, J., Girard, S., Peralta, E. & Clapham, D. Spiral calcium wave propagation and annihilation in *Xenopus laevis* oocytes. *Science* **252**, 123–126 (1991).
6. Gorelova, N. A. & Bures, J. Spiral waves of spreading depression in the isolated chicken retina. *J. Neurobiol.* **14**, 353–363 (1983).
7. Davidenko, J. M., Pertsov, A. V., Salomonsz, R., Baxter, W. & Jalife, J. Stationary and drifting spiral waves of excitation in isolated cardiac muscle. *Nature* **355**, 349–351 (1992).
8. Gray, R. A., Pertsov, A. M. & Jalife, J. Spatial and temporal organization during cardiac fibrillation. *Nature* **392**, 75–78 (1998).
9. Kim, M. *et al.* Controlling chemical turbulence by global delayed feedback: pattern formation in catalytic CO oxidation on Pt(110). *Science* **292**, 1357–1360 (2001).
10. Tung, C. K. & Chan, C. K. Dynamics of spiral waves under phase feedback control in a Belousov–Zhabotinsky reaction. *Phys. Rev. Lett.* **89**, 248302 (2002).
11. Saran, S. *et al.* cAMP signaling in *Dictyostelium*—Complexity of cAMP synthesis, degradation and detection. *J. Muscle Res. Cell Motil.* **23**, 793–802 (2002).
12. Iranfar, N., Fuller, D. & Loomis, W. F. Genome-wide expression analyses of gene regulation during early development of *Dictyostelium discoideum*. *Euk. Cell* **2**, 664–670 (2003).
13. Gross, J. D., Peacey, M. J. & Trevan, D. J. Signal emission and signal propagation during early aggregation in *Dictyostelium discoideum*. *J. Cell Sci.* **22**, 645–656 (1976).

14. Devreotes, P. N., Potel, M. J. & MacKay, S. A. Quantitative analysis of cyclic AMP waves mediating aggregation in *Dictyostelium discoideum*. *Dev. Biol.* **96**, 405–415 (1983).
15. Torrence, C. & Compo, C. P. A practical guide to wavelet analysis. *Bull. Am. Meteorol. Soc.* **79**, 61–78 (1998).
16. Thomason, P. A. *et al.* An intersection of the cAMP/PKA and two-component signal transduction systems in *Dictyostelium*. *EMBO J.* **17**, 2838–2845 (1998).
17. Wessels, D. J. *et al.* The internal phosphodiesterase RegA is essential for the suppression of lateral pseudopods during *Dictyostelium* chemotaxis. *Mol. Biol. Cell* **11**, 2803–2820 (2000).
18. Harwood, A. J. *et al.* Culmination in *Dictyostelium* is regulated by the cAMP-dependent protein kinase. *Cell* **69**, 615–624 (1992).
19. Patel, H. *et al.* A temperature-sensitive adenylyl cyclase mutant of *Dictyostelium*. *EMBO J.* **19**, 2247–2256 (2000).
20. Segall, J. E. *et al.* A MAP kinase necessary for receptor-mediated activation of adenylyl cyclase in *Dictyostelium*. *J. Cell Biol.* **128**, 405–413 (1995).
21. Maeda, M. *et al.* Periodic signaling controlled by an oscillatory circuit that includes protein kinases ERK2 and PKA. *Science* **304**, 875–878 (2004).
22. Sugang, R., Weijer, C. J., Siegfert, F., Franke, J. & Kessin, R. H. Null mutations of the *Dictyostelium* cyclic nucleotide phosphodiesterase gene block chemotactic cell movement in developing aggregates. *Dev. Biol.* **192**, 181–192 (1997).
23. Laub, M. T. & Loomis, W. F. A molecular network that produces spontaneous oscillations in excitable cells of *Dictyostelium*. *Mol. Biol. Cell* **9**, 3521–3532 (1998).
24. Martiel, J.-L. & Goldbeter, A. A model based on receptor desensitization for cyclic AMP signaling in *Dictyostelium* cells. *Biophys. J.* **52**, 807–828 (1987).
25. Levine, H., Aranson, I., Tsimring, L. & Truong, T. V. Positive genetic feedback governs cAMP spiral wave formation in *Dictyostelium*. *Proc. Natl Acad. Sci. USA* **93**, 6382–6386 (1996).
26. Lee, K. J., Goldstein, R. E. & Cox, E. C. Resetting wave forms in *Dictyostelium* territories. *Phys. Rev. Lett.* **87**, 066101 (2001).
27. Fey, P., Compton, K. & Cox, E. C. Green fluorescent protein production in the cellular slime molds *Polysphondylium pallidum* and *Dictyostelium discoideum*. *Gene* **165**, 127–130 (1995).
28. Simon, M. N., Pelegri, O., Veron, M. & Kay, R. R. Mutation of protein kinase-A causes heterochronic development of *Dictyostelium*. *Nature* **356**, 171–172 (1992).

Supplementary Information accompanies the paper on www.nature.com/nature.

Acknowledgements The authors thank X. J. Guan and other members of the Cox laboratory for assistance, J. E. Segall for HS174 and HS174-A1, A. Kuspa for AK1045, H. Kuwayama for the *pkaC*-null strain, A. Harwood for the *pka*-Rm and *Re* expression vectors, M. Veron for the anti-PKA-R monoclonal antibody, J. T. Bonner and M. Sano for discussions and K. McQuade, R. Segev and G. Weeks for detailed comments on our earlier manuscript. The authors also thank the *Dictyostelium* Stock Centre for providing HTY506. P.A.T. received a fellowship from the Wellcome Trust. This work was supported by a grant to E.C.C. from the NIH.

Competing interests statement The authors declare that they have no competing financial interests.

Correspondence and requests for materials should be addressed to E.C.C. (ecox@princeton.edu) or S.S. (ssawai@molbio.princeton.edu).

A pentatricopeptide repeat protein is essential for RNA editing in chloroplasts

Emi Kotera¹, Masao Tasaka¹ & Toshiharu Shikanai^{1,2}

¹Graduate School of Biological Sciences, Nara Institute of Science and Technology, Ikoma, Nara 630-0101, Japan

²Graduate School of Agriculture, Kyushu University, Fukuoka 812-8581, Japan

RNA editing is a process of RNA maturation involved in the insertion, deletion or modification of nucleotides¹. In organellar transcripts of higher plants, specific cytidine residues are converted into uridine residues. In many cases, editing results in the restoration of conserved amino acid residues, a process that is essential for protein function in plastids^{2,3}. Despite the technical breakthrough in establishing systems *in vivo*⁴ and *in vitro*⁵ for analysing RNA editing, its machinery still remains to be identified in higher plants. Here we introduce a genetic approach and report the discovery of a gene responsible for the specific RNA editing event in the chloroplast.

The chloroplast NA(P)DH dehydrogenase (NDH) complex is involved in cyclic electron flow around photosystem I (refs 6, 7). Eleven subunits of the complex are encoded in the chloroplast genome in higher plants (*ndhA-ndhK*). We developed a screening system to identify mutants with impaired NDH activity in *Arabidopsis thaliana*⁸. In doing so, we were also able to identify the mutants affected in RNA editing.

Four *crr4* (chlororespiratory reduction) mutants, *crr4-1* to *crr4-4*, were identified by their lack of post-illumination increase in chlorophyll fluorescence (Fig. 1a, b) using a modified system of fluorescence imaging^{8,9}. This transient rise in fluorescence is due to the dark reduction of the plastoquinone pool, which is dependent on NDH activity (Fig. 1a). Whereas the fluorescence rise was undetectable in three alleles, *crr4-1*, *crr4-2* and *crr4-3*, a trace of activity was detected in *crr4-4* (Fig. 1b). Genetic analysis established that *crr4* mutations are recessive, with four mutations on a single gene (data not shown). With the exception of the lack of NDH activity, *crr4* mutants did not display any phenotype in photo-

synthetic electron transport, which is consistent with other mutants specifically impaired in NDH activity^{6,8}. Western analysis using an antibody against NdhH showed a decrease in the level of the NDH complex (Fig. 1c). NdhH was almost undetectable in the two alleles *crr4-1* and *crr4-3*, and a trace level was detected in *crr4-2*. The NdhH level was mildly affected in *crr4-4* (25–50%).

To identify a gene affected in *crr4* mutants, *crr4-1* was crossed with the polymorphic wild type (*Landsberg erecta*). Analysis of 243 F₂ plants specified a 305-kilobase region on chromosome 2. Because the *crr4* phenotype was specific to the chloroplast NDH level, candidate genes encoding the putative chloroplast targeting signal were sequenced. Finally, mutations were found in gene At2g45350 in all *crr4* alleles. To confirm the *crr4* phenotype was caused by mutations in At2g45350, the wild-type genomic sequence of At2g45350 was introduced into *crr4-1* and *crr4-2*. The transformation fully complemented the transient increase in chlorophyll fluorescence (Fig. 1b) and NdhH level (Fig. 1c).

CRR4 (At2g45350) was not interrupted by any intron and encoded a putative protein with 606 amino acids. The program ChloroP 1.1 predicted that the first 45 amino acids were the target signal to the plastid (Fig. 2a). CRR4 is a member of a PCMP (for plant combinatorial and modular protein) family¹⁰. On the basis of structural and functional similarities, the PCMP family is included in the pentatricopeptide repeat (PPR) family^{8,11}, as the PLS sub-family¹². The PPR family consists of about 450 genes in the *Arabidopsis* genome. CRR4 contains 11 PPR motifs consisting of a degenerate 35-amino-acid unit (Fig. 2b). Overlapping the array of PPR motifs, the central region of CRR4 comprises the reiteration of three motifs, A, B and C, in the order 3 × (B-C-A), followed by two

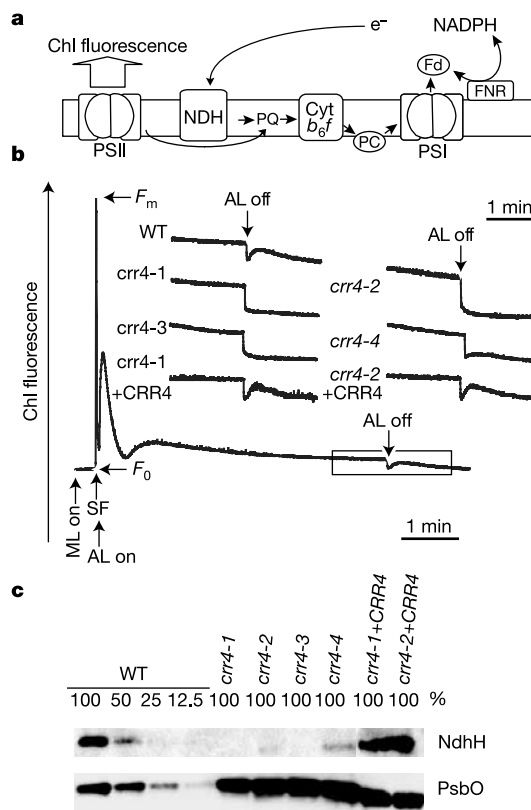


Figure 1 Characterization of *crr4* mutants. **a**, Schematic model of NDH function. The NDH complex functions in electron transport from the stromal reducing pool, NADPH and reduced ferredoxin (Fd), to the plastoquinone pool (PQ). PQ reduction in the dark depends on NDH activity and is detected in the transient rise of chlorophyll fluorescence after illumination with actinic light (AL)⁶. PSI, photosystem I; PSII, photosystem II. **b**, Analysis of the transient increase in chlorophyll fluorescence after turning off AL. The bottom curve indicates a typical trace of chlorophyll fluorescence in the wild type (WT). Leaves were exposed to AL (50 $\mu\text{mol photons m}^{-2} \text{s}^{-1}$) for 5 min. AL was turned off and the subsequent transient rise in fluorescence ascribed to NDH activity was monitored by chlorophyll fluorimetry. Insets are magnified traces from the boxed area. *crr4-X*+CRR4, *crr4* alleles transformed by the wild-type genomic CRR4 sequence. ML, measuring light; SF, saturating flash. **c**, Immunoblot analysis of thylakoid proteins. Immunodetection of an NDH subunit (NdhH) and photosystem II (PsbO). The lanes were loaded with proteins corresponding to 0.5 μg of chlorophyll for PsbO and tenfold the proteins for NdhH (100%) and a series of dilutions as indicated.

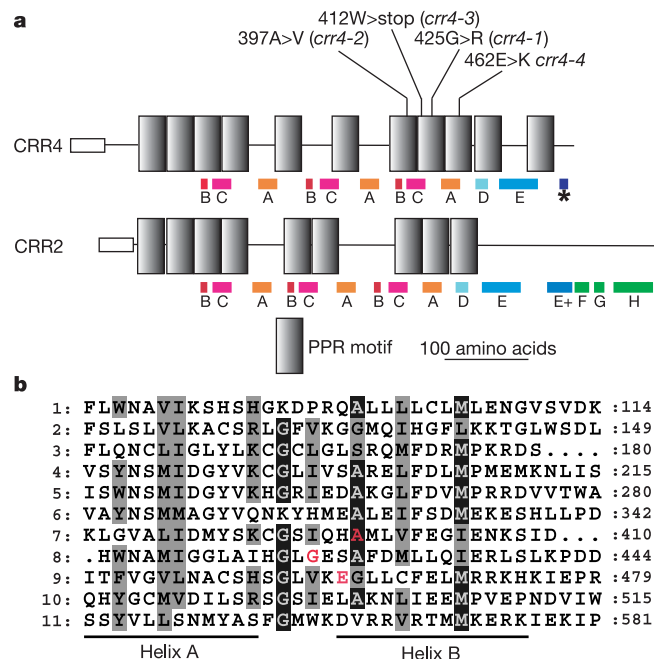


Figure 2 Structure of CRR4. **a**, Schematic alignment of CRR4 and CRR2. The relationship between PPR motifs (boxed) and PCMP motifs (bars A–H) is shown. PCMP motifs are labelled on the basis of the original assignment¹⁰ except for the E+ motif¹². The sites of mutation in four *crr4* alleles are indicated. White boxes are the putative plastid targeting signals. The C-terminal 15-amino-acid motif is indicated by a bar labelled with an asterisk. **b**, Alignment of eleven PPR motifs present in CRR4. Amino acids conserved more than 60% are boxed in black. Conserved similar amino acids are shaded. The points of amino acid alteration in three alleles are highlighted by red letters. A pair of antiparallel α -helices, predicted from the similarity with TPR motif¹¹, is shown by underlines.

carboxy-terminal motifs, D and E (Fig. 2a). CRR4 belongs to the PCMP-E subfamily (E subgroup) lacking the three motifs F, G and H, which are conserved in the PCMP-H subfamily (DYW subgroup)^{10,12}. The final PPR motif is followed by an unknown motif consisting of 15 amino acids (PGCSxI/VExxGxV/IHxV), which is highly conserved in some PPR proteins. Although this motif is apparently related to E and E+ motifs¹², it is well conserved only in some members.

All the amino acid alterations present in three alleles of *crr4* are located on the third iteration of the B-C-A motif, indicating that the region might be essential for CRR4 function (Fig. 2a). They are located close to each other in the predicted three-dimensional structure of PPR (antiparallel α -helices)¹¹ (Fig. 2b). In contrast, *crr4-3* has a nonsense mutation resulting in truncation of the protein after the seventh PPR motif, suggesting that *crr4-3* is a null allele.

Some members of the PPR family are involved in RNA processing, stabilization and translation in the chloroplasts^{8,13–16} and mitochondria of yeast¹⁷, *Neurospora*¹⁸ and higher plants^{19–21}. This fact indicates that RNA maturation of the chloroplast *ndh* gene(s) might be impaired in *crr4*. However, no alteration in the size or level of the transcripts derived from the 11 *ndh* genes was found (data not shown).

Although the *crr4* phenotype was not detected in the RNA blot analysis, it is still possible that *crr4* is defective in RNA editing or translation. To study the former possibility, products of polymerase chain reaction with reverse transcription (RT-PCR) were sequenced directly (Fig. 3). Among 11 *Arabidopsis ndh* genes, *ndhA*, *ndhB*, *ndhD* and *ndhF* contained one, eight, four and one editing sites, respectively²². In tobacco *ndhD*, the translation initiation codon is encoded by ACG in the genome, which is edited to AUG by RNA editing²³. The efficiency of this editing depends on the tissue and the age of the leaf²⁴. Figure 3a shows the result of the direct sequencing of RT-PCR products around the *ndhD* translation initiation site. Consistent with the results in tobacco, the translation initiation codon ACG was partly edited in *Arabidopsis* leaves. In contrast, this

editing site remained unedited in the transcripts from *crr4* mutants. Figure 3b shows the extent of RNA editing after semiquantitative analysis with a restriction enzyme that digests only complementary DNA derived from edited RNA molecules. RT-PCR products subjected to the analysis covered the region from –38 to +254 (taking the initiation codon of *ndhD* as +1), which is totally included in the monocistronic *ndhD* transcripts (Supplementary Fig. 1). A similar result was obtained with longer RT-PCR products ranging from –38 to +996 as a probe (data not shown). About 50% of RNA molecules were edited in mature leaves of *Arabidopsis*, although the edited molecules were hardly detectable in *crr4* (Fig. 3b).

To evaluate the efficiency of RNA editing more directly, the editing site within the cDNA was amplified by PCR and cloned in *Escherichia coli*. About 100 independent clones were analysed by digestion with *NlaIII* to detect cDNA originating from the edited RNA molecules (Supplementary Table 1). Consistent with the estimation of semiquantitative analysis, 48% of molecules were edited in the wild type. In contrast, no molecules were edited in *crr4-3*, supporting the idea that *crr4-3* is a null allele. Detection of a single edited clone out of 111 clones examined indicates that the missense allele of *crr4-1* is unlikely to be null. It is also possible that the low level of this editing is due to an inefficient cross-reaction by an alternative PPR protein. Three edited clones of the 106 clones examined were detected in *crr4-4*, which is consistent with our conclusion that *crr4-4* is a weak allele.

Although the RNA editing in the initiation codon of *ndhD* was impaired in *crr4*, all known editing sites present in *ndhB*, *ndhD*,

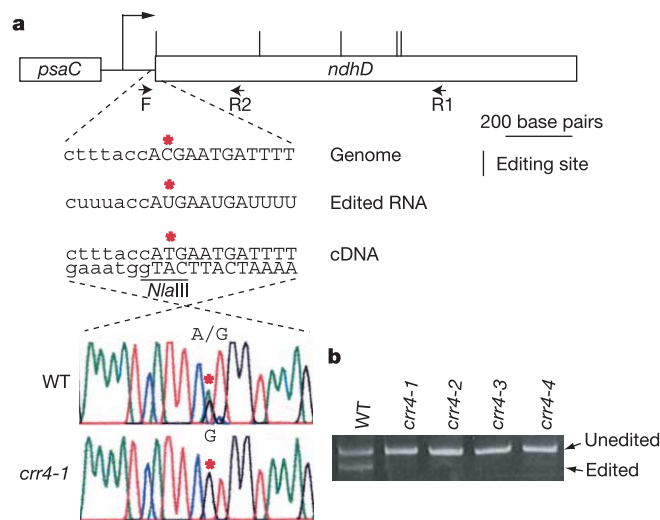


Figure 3 Analysis of RNA editing in the *ndhD* initiation codon. **a**, Direct sequencing of RT-PCR products containing the *ndhD* initiation codon. The *psaC* and *ndhD* region is shown schematically. RNA editing sites are indicated. The restriction enzyme *NlaIII* cleaves cDNA derived from edited molecules. The editing site is so distal in the transcripts that cDNA was sequenced on the complementary strand. **b**, Semi-quantitative analysis of the extent of RNA editing. RT-PCR products were digested with *NlaIII*. Fragments originating from edited and unedited RNA molecules are indicated. WT, wild type.

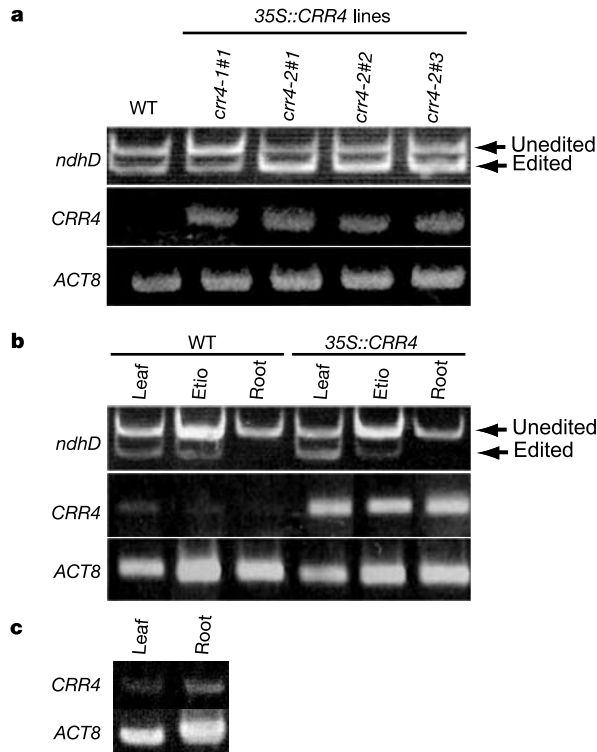


Figure 4 Relationship between the *CRR4* expression level and the extent of RNA editing. **a**, Semi-quantitative analysis of RNA editing in the *ndhD* initiation codon in lines overexpressing *CRR4*. The accumulation of *CRR4* RNA was estimated by RT-PCR. *ACT8* is a control for RT-PCR. WT, wild type. **b**, Semi-quantitative analysis of RNA editing in the *ndhD* initiation codon in leaf, etiolated seedling (etio) and root cells. RNA was extracted from the *35S::CRR4* line (*crr4-1#1*), as well as from the wild type. **c**, *CRR4* is expressed in both roots and leaves. Cycles of RT-PCR were 26 in **b** and 29 in **c**.

ndhF, *atpF*, *psbE*, *rpoB* and *rps14* (ref. 22) were processed, as they were in the wild type (Supplementary Table 2). We conclude that *crr4* is specifically defective in the RNA editing that creates an initiation codon of *ndhD*.

In *Arabidopsis*, *ndhD* is polycistronically transcribed with the upstream gene *psaC* (Fig. 3a). Post-transcriptional RNA processing generates monocistronic RNA species that are translatable^{24,25}. Because the *ndhD* initiation codon is closely located to these RNA cleavage sites, it is possible that *crr4* is defective in these RNA cleavages and is consequently defective in the RNA editing. To eliminate this possibility, RNA processing between *psaC* and *ndhD* was analysed with the RNase protection assay in *crr4* and the wild type. Neither the 3' termini of *psaC* nor the 5' termini of *ndhD* were altered in *crr4* (Supplementary Fig. 1).

Our results are inconsistent with a previous report indicating that the *ndhD* initiation codon was not edited in *Arabidopsis*²². We were also unable to find the reported RNA editing in *ndhA* (Supplementary Table 2). Some RNA editing processes might depend on tissue and culture conditions²⁴. However, it is unlikely that the *crr4* defect in the RNA editing is caused secondarily by a physiological change, because the absence of the NDH complex hardly affected the chloroplast function under the culture conditions used. It is also unlikely that the absence of the NDH complex directly affected the RNA editing, because the editing was unaffected in *pgr3-3*, in which accumulation of the NDH complex was impaired¹⁵. We conclude that *crr4* is directly defective in the RNA editing, rather than indirectly through the other defects affecting the editing process.

A defect in RNA editing in the *ndhD* initiation codon drastically impaired the accumulation of the NDH complex (Fig. 1c). Despite the physiological significance, this editing was partial, even in the wild type. The extent of the editing was dependent on the developmental and environmental conditions²⁴. Because CRR4 is specifically involved in this RNA editing, the CRR4 level might be a determinant of the extent of editing and consequently of *ndhD* translation. To study this possibility, we analysed the extent of *ndhD* editing in the *crr4* lines in which CRR4 was overexpressed under the control of the 35S promoter. Introduction of the 35S::CRR4 construct to *crr4-1* and *crr4-2* restored the RNA editing activity (Fig. 4a). Despite the upregulation of CRR4 expression at least in the RNA level (Fig. 4a, CRR4), the extent of RNA editing did not increase in comparison with the wild-type level. This result suggests that the efficiency of RNA editing is not simply limited by CRR4 expression.

To study further the factor limiting the editing efficiency, RNA was prepared from roots in which the NDH complex did not accumulate²⁶. Consistent with the result in tobacco²⁶ was our observation that the *ndhD* initiation codon was unedited in *Arabidopsis* roots (Fig. 4b). However, RT-PCR detected the *crr4* transcript in both roots and leaves (Fig. 4b, c). Furthermore, overexpression of CRR4 did not promote the RNA editing of the *ndhD* initiation codon in roots. This observation is consistent with the fact that 15 sites on the transcript from *ndh* genes are edited inefficiently in tobacco roots²⁶. An unknown factor involved in at least 15 editing sites seemed to limit RNA editing. RNA was also extracted from etiolated seedlings. Consistent with the previous report²⁷, the *ndhD* initiation codon was partly edited in etiolated seedlings. Overexpression of CRR4 did not increase the editing efficiency in etiolated seedlings. Combining this result with results in the leaves and roots, we eliminate the simplest model in which the CRR4 level determines the editing efficiency. CRR4 might be part of a complex (see below) and subunits of the complex other than CRR4 might be rate-limiting. It is also possible that CRR4 is modified to facilitate RNA recognition.

Accumulating information suggests that the PPR protein recognizes the specific RNA for maturation^{8,13–21}. This is consistent with our finding that the *crr4* defect in RNA editing is specific to the *ndhD* initiation codon. Figure 2a shows the alignment of two PPR/PCMP

family members, CRR4 and CRR2. CRR2 functions in the intergenic RNA cleavage between *rps7* and *ndhB*, which is essential for *ndhB* translation⁸. CRR2 belongs to the PCMP-H subfamily (DYW subgroup) and has well-conserved C-terminal motifs, F, G and H, which are missing from CRR4. Although CRR2 and CRR4 are involved in different processes of RNA maturation, they are closely related in structure. All the motifs included in CRR4 are present in CRR2 except the C-terminal 15-amino-acid motif, which is less conserved in CRR2 (Fig. 2). However, this 15-amino-acid motif is too small to explain on its own the activity that deaminates a specific cytidine residue, the reaction that is required for the RNA editing process. We propose that a PPR protein functions as a factor for recognizing the target RNA and facilitates the access of a general factor containing cytidine deaminase activity. This simplest model is partly analogous to the editing system of the human APOB100 locus, in which an RNA-binding protein, ACF (apobec-1 complementation factor), facilitates the access of a cytidine deaminase, APOBEC1, to the editing site²⁸. However, a multi-component system is not consistent with a recent report suggesting that both site determination and C → U conversion are governed by a single factor in the chloroplast²⁹. Our findings seem to support a multi-component system, in which the PPR protein might interact with the target RNA to recruit the editing machinery. Considering the extraordinarily large size of the PPR family, it is reasonable to expect that PPR proteins are involved in RNA editing in organelles, by which 19 and 441 sites are processed in the chloroplast and the mitochondrion of *Arabidopsis*, respectively^{22,30}. By incorporating PPR proteins in the RNA editing machinery, higher plants can manage several hundreds of RNA editing events in organelles. □

Methods

Plant materials and growth conditions

The accession of *crr4* mutants is Columbia *gl1*. All alleles were derived from mutagenesis with ethylmethane sulphonate. Mutants were selected as described previously⁸. Plants were grown in Metromix potting soil under controlled conditions (light intensity of 50 $\mu\text{mol photons m}^{-2} \text{s}^{-1}$, 16 h light/8 h dark cycle at 23 °C).

Chlorophyll fluorescence analysis

Chlorophyll fluorescence was measured with a MINI-PAM portable chlorophyll fluorimeter (Walz). The transient increase in chlorophyll fluorescence after turning off the actinic light (50 $\mu\text{mol photons m}^{-2} \text{s}^{-1}$ for 5 min) was monitored as described⁶.

Map-based cloning

The *crr4* mutations were mapped with molecular markers based on cleaved amplified polymorphic sequence. *RsaI* and *HinfI* were used for markers T13E15 and F418, respectively (Supplementary Table 3). Genomic sequences of At2g45350 from the wild-type and *crr4* alleles were amplified by PCR with ExTaq DNA polymerase (Takara). The products were sequenced directly with a dye terminator cycle sequencing kit and an ABI PRISM 3100 sequencer.

For complementation of the *crr4* mutations, the wild-type genomic sequence from –1860 to +2240 (taking the translation initiation site as +1) was amplified by PCR and cloned in pBIN19. For CRR4 overexpression, the wild-type genomic sequence from –135 to +2240 was cloned in pBI121. The resulting plasmids were introduced into *Agrobacterium tumefaciens* MP90 and then into *crr4-1* and *crr4-2*.

Immunoblot analysis

Thylakoid proteins were isolated from leaves of 4-week-old plants, fractionated by 12.5% SDS-PAGE, and transferred to poly(vinylidene difluoride) membranes. NdhH and PsbO were detected with the enhanced chemiluminescence detection kit (Amersham).

Analysis of RNA editing

Total RNA was isolated from rosette leaves by using an RNAeasy plant mini kit (Qiagen) and treated with DNase (Invitrogen) to remove any trace of the genome DNA. DNA-free RNA (1 μg) was reverse transcribed with random hexamers. Sequences including the editing sites were amplified by PCR with primers in the combinations *ndhD-F/ndhD-R1* or *ndhD-F/ndhD-R2* (Supplementary Table 3). The RT-PCR products were sequenced directly. For semiquantitative analysis of the initiation codon of *ndhD*, the RT-PCR products were digested with *NlaIII* and analysed on a 10% polyacrylamide gel.

Received 17 August; accepted 26 November 2004; doi:10.1038/nature03229.

1. Smit, H. C., Gott, J. M. & Hanson, M. R. A guide to RNA editing. *RNA* 3, 1105–1123 (1997).
2. Bock, R., Kössel, H. & Maliga, P. Introduction of a heterologous editing site into the tobacco

- plastid genome: the lack of RNA editing leads to a mutant phenotype. *EMBO J.* **13**, 4623–4628 (1994).
3. Sasaki, Y., Kozaki, A., Ohmori, A., Iguchi, H. & Nagano, Y. Chloroplast RNA editing required for functional acetyl-CoA carboxylase in plants. *J. Biol. Chem.* **276**, 3937–3940 (2000).
 4. Chaudhuri, S., Carrer, H. & Maliga, P. Site-specific factor involved in the editing of the *psbL* mRNA in tobacco plastids. *EMBO J.* **14**, 2951–2957 (1995).
 5. Hirose, T. & Sugiura, M. Involvement of a site-specific trans-acting factor and a common RNA-binding protein in the editing of chloroplast mRNAs: development of a chloroplast *in vitro* RNA editing system. *EMBO J.* **20**, 1144–1152 (2001).
 6. Shikanai, T. *et al.* Directed disruption of the tobacco *ndhB* gene impairs cyclic electron flow around photosystem I. *Proc. Natl Acad. Sci. USA* **95**, 9705–9709 (1998).
 7. Munekage, Y. *et al.* Cyclic electron flow around photosystem I is essential for photosynthesis. *Nature* **429**, 579–582 (2004).
 8. Hashimoto, M., Endo, T., Peltier, G., Tasaka, M. & Shikanai, T. A nucleus-encoded factor, CRR2, is essential for the expression of chloroplast *ndhB* in *Arabidopsis*. *Plant J.* **36**, 541–549 (2003).
 9. Shikanai, T., Munekage, Y., Shimizu, K., Endo, T. & Hashimoto, T. Identification and characterization of *Arabidopsis* mutants with reduced quenching of chlorophyll fluorescence. *Plant Cell Physiol.* **40**, 1134–1142 (1999).
 10. Aubourg, S., Boudet, N., Kreis, M. & Lechamy, A. In *Arabidopsis thaliana*, 1% of the genome codes for a novel protein family unique to plants. *Plant Mol. Biol.* **42**, 603–613 (2000).
 11. Small, I. D. & Peeters, N. The PPR motif—a TPR-related motif prevalent in plant organellar proteins. *Trends Biochem. Sci.* **25**, 46–47 (2000).
 12. Lurin, C. *et al.* Genome-wide analysis of *Arabidopsis* pentatricopeptide repeat proteins reveals their essential role in organelle biogenesis. *Plant Cell* **16**, 2089–2103.
 13. Fisk, D. G., Walker, M. B. & Barkan, A. Molecular cloning of the maize gene *crp1* reveals similarity between regulators of mitochondrial and chloroplast gene expression. *EMBO J.* **18**, 2621–2630 (1999).
 14. Meierhoff, K., Felder, S., Nakamura, T., Bechtold, N. & Schuster, G. HCF152, an *Arabidopsis* RNA binding pentatricopeptide repeat protein involved in the processing of chloroplast *psbH-psbI-psbH-petD* RNAs. *Plant Cell* **15**, 1480–1495 (2003).
 15. Yamazaki, H., Tasaka, M. & Shikanai, T. PPR motifs of the nucleus-encoded factor, PGR3, function in the selective and distinct steps of chloroplast gene expression in *Arabidopsis*. *Plant J.* **38**, 152–163 (2004).
 16. Williams, P. M. & Barkan, A. A chloroplast-localized PPR protein required for plastid ribosome accumulation. *Plant J.* **36**, 675–686 (2003).
 17. Manthey, M. G. & McEwen, J. E. The product of the nuclear gene *PET309* is required for translation of mature mRNA and stability or production of intron-containing RNAs derived from the mitochondrial *COX1* locus of *Saccharomyces cerevisiae*. *EMBO J.* **14**, 4031–4043 (1995).
 18. Coffin, J. W., Dhillon, R., Ritzel, R. G. & Nargang, E. E. The *Neurospora crassa* *cya-5* nuclear gene encodes a protein with a region of homology to the *Saccharomyces cerevisiae* PET309 protein and is required in a post-transcriptional step for the expression of the mitochondrially encoded COXI protein. *Curr. Genet.* **32**, 273–280 (1997).
 19. Bentolila, S., Alfonso, A. A. & Hanson, M. R. A pentatricopeptide repeat-containing gene restores fertility to cytoplasmic male-sterile plants. *Proc. Natl Acad. Sci. USA* **99**, 10887–10892 (2002).
 20. Kazama, T. & Toriyama, K. A pentatricopeptide repeat-containing gene that promotes the processing of aberrant *atp6* RNA of cytoplasmic male-sterile rice. *FEBS Lett.* **544**, 99–102 (2003).
 21. Koizuka, N. *et al.* Genetic characterization of a pentatricopeptide repeat protein gene, *orf687*, that restores fertility in the cytoplasmic male-sterile Kosenada radish. *Plant J.* **34**, 407–415 (2003).
 22. Lutz, K. A. & Maliga, P. Lack of conservation of editing sites in mRNAs that encode subunits of the NAD(P)H dehydrogenase complex in plastids and mitochondria of *Arabidopsis thaliana*. *Curr. Genet.* **40**, 214–219 (2001).
 23. Neckermann, K., Zeltz, P., Igloi, G. L., Kössel, H. & Maier, R. M. The role of RNA editing in conservation of start codons in chloroplast genomes. *Gene* **146**, 177–182 (1994).
 24. Hirose, T. & Sugiura, M. Both RNA editing and RNA cleavage are required for translation of tobacco chloroplast *ndhD* mRNA: a possible regulatory mechanism for the expression of a chloroplast operon consisting of functionally unrelated genes. *EMBO J.* **16**, 6804–6811 (1997).
 25. del Campo, E. M., Sabater, B. & Martin, M. Post-transcriptional control of chloroplast gene expression. Accumulation of stable *psaC* mRNA is due to downstream RNA cleavages in the *ndhD* gene. *J. Biol. Chem.* **277**, 36457–36464 (2002).
 26. Chateigner-Boutin, A.-L. & Hanson, M. R. Developmental co-variation of RNA editing extent of plastid editing sites exhibiting similar *cis*-elements. *Nucleic Acids Res.* **31**, 2586–2594 (2003).
 27. Guera, A., Nova, P. G. & Sabater, B. Identification of the Ndh (NAD(P)H-plastoquinone-oxidoreductase) complex in etioplast membranes of barley: changes during photomorphogenesis of chloroplasts. *Plant Cell Physiol.* **41**, 49–59 (2000).
 28. Keegan, L. P., Gallo, A. & O'Connell, M. A. The many roles of an RNA editor. *Nature Rev. Genet.* **2**, 869–878 (2001).
 29. Miyamoto, T., Obokata, J. & Sugiura, M. A site-specific factor interacts directly with its cognate RNA editing site in chloroplast transcripts. *Proc. Natl Acad. Sci. USA* **101**, 48–52 (2004).
 30. Glegé, P. & Brennicke, A. RNA editing in *Arabidopsis* mitochondria effects 441 C to U changes in ORFs. *Proc. Natl Acad. Sci. USA* **96**, 15324–15329 (1999).

Supplementary Information accompanies the paper on www.nature.com/nature.

Acknowledgements We thank M. Miyata and E. Habe for their technical assistance; T. Endo, G. Peltier and the late A. Watanabe for antibodies; and J. Obokata for suggestions. T. Shikanai was supported by grants from JSPS and Ministry of Education, Sports, Culture, Science and Technology, Japan.

Competing interests statement The authors declare that they have no competing financial interests.

Correspondence and requests for materials should be addressed to T.S. (shikanai@agr.kyushu-u.ac.jp).

Molecular dynamics of cyclically contracting insect flight muscle *in vivo*

Michael Dickinson¹, Gerrie Farman^{2*}, Mark Frye^{1*}, Tanya Bekyarova², David Gore², David Maughan³ & Thomas Irving²

¹Department of Bioengineering, California Institute of Technology, Pasadena, California 91125, USA

²BioCAT and CSRR, Department BCPS, Illinois Institute of Technology, Chicago, Illinois 60616, USA

³Department of Molecular Physiology and Biophysics, University of Vermont, Burlington, Vermont 05405, USA

* Present addresses: Department of Physiology and Biophysics, University of Illinois at Chicago, Chicago, Illinois 60612, USA (G.F.); Department of Physiological Science, University of California, Los Angeles, Los Angeles, California 90095, USA (M.F.).

Flight in insects—which constitute the largest group of species in the animal kingdom—is powered by specialized muscles located within the thorax. In most insects each contraction is triggered not by a motor neuron spike but by mechanical stretch imposed by antagonistic muscles¹. Whereas ‘stretch activation’ and its reciprocal phenomenon ‘shortening deactivation’ are observed to varying extents in all striated muscles, both are particularly prominent in the indirect flight muscles of insects¹. Here we show changes in thick-filament structure and actin–myosin interactions in living, flying *Drosophila* with the use of synchrotron small-angle X-ray diffraction. To elicit stable flight behaviour and permit the capture of images at specific phases within the 5-ms wingbeat cycle, we tethered flies within a visual flight simulator². We recorded images of 340 μ s duration every 625 μ s to create an eight-frame diffraction movie, with each frame reflecting the instantaneous structure of the contractile apparatus. These time-resolved measurements of molecular-level structure provide new insight into the unique ability of insect flight muscle to generate elevated power at high frequency.

The asynchronous indirect flight muscles (IFMs) of insects have served as a valuable model system for general structural studies of muscle because of the well-ordered arrangement of the contractile lattice. X-ray diffraction studies (for example refs 3–6) and electron microscopy studies (for example refs 3, 5, 7, 8) of glycerol-extracted fibres from the giant waterbug, *Lethocerus* sp., in combination with crystallographic information, have produced a high-resolution view of actin–myosin interactions^{6,8,9}. Previous synchrotron X-ray diffraction studies of *Drosophila*¹⁰ demonstrated the feasibility of obtaining patterns from intact IFMs in live flies. Because of the similarity of the IFM structures, the extensive literature from preparations of *Lethocerus* IFM *in vitro* can be used as a framework for interpreting changes in X-ray diffraction patterns obtained from living, flying *Drosophila*. Here we optically tracked the motion of the beating wings in real time and triggered the X-ray exposure at specific phases of the wingbeat cycle (Fig. 1). The beam was focused on the dorsal longitudinal muscles (DLMs), the IFMs that contract during the downstroke. The tethered animal was surrounded by an electronic display of vertical stripes, the motion of which was controlled by the insect’s own wing motion under closed-loop conditions². This arrangement maintained a stable wingbeat frequency, permitting accurate shuttering at precise phase points within the contraction cycle. By progressing through the wingbeat, we reconstructed a movie of X-ray diffraction images from the DLMs. Figure 2 shows examples of such images obtained from a single *Drosophila metleri* (the full sequence is shown as an animation in Supplementary Video 1). The cyclical changes in the intensities and positions of the major reflections reflect the cyclical binding,

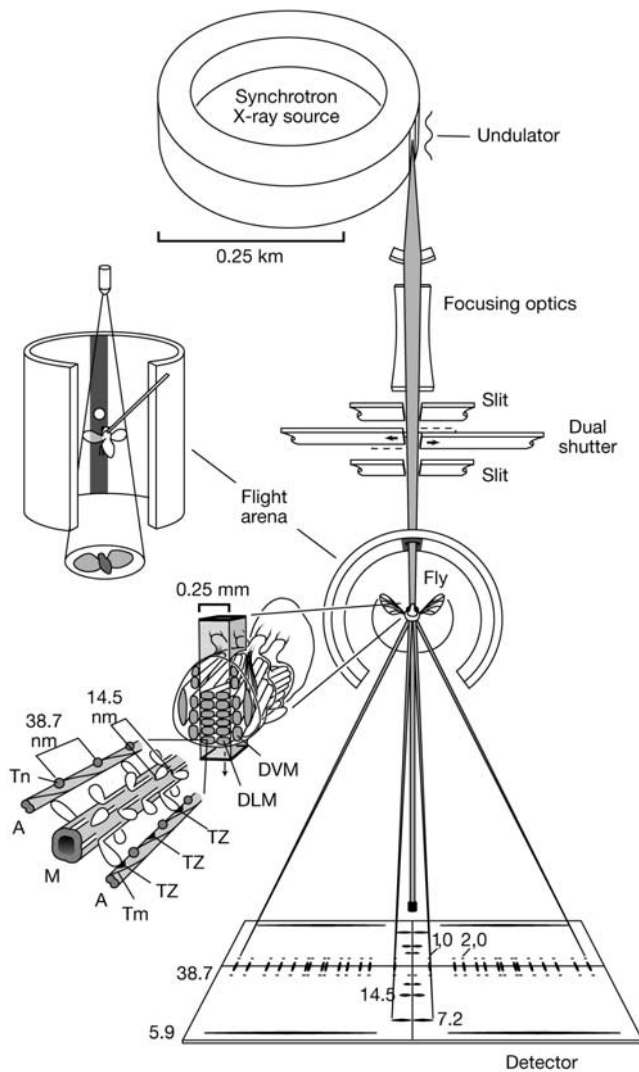


Figure 1 Experimental arrangement. The position of the fly is shown relative to the flight arena and the X-ray optical path. The locations of the DLMs and DVMs with respect to the incident X-ray beam are indicated in the expanded view. Periodic structures of the filament lattice that give rise to the features seen in diffraction patterns are also indicated (Tn, troponin complex; Tm, tropomyosin). Note the axial rise at 14.5 nm between myosin (M) heads along the thick filaments and the location of target zones (TZ) for the binding of myosin heads to actin (A) halfway between troponins, which are spaced every 38.7 nm along the thin filament.

tilting and release of myosin heads on actin target zones during the wingbeat (Fig. 3).

In vertebrate muscle, the intensity of the 14.5-nm meridional reflection has been shown to arise mainly from the axial projection of the density of myosin heads onto the thick-filament backbone^{11–13}, whereas higher-order reflections, such as that at 7.2 nm, are thought to arise from periodic features within the backbone itself^{14,15}. Therefore, changes in the position of the 7.2-nm reflection provide a measure of thick-filament backbone strain, whereas motion of the 14.5-nm reflection is caused by changes in the position of the centre of mass of the myosin heads in series with the backbone strain. Figure 3a shows the relationship of the spacing of the 7.2-nm and 14.5-nm meridional reflections relative to rest throughout the upstroke–downstroke cycle of the wings. Both spacings are tightly coupled to the lengthening–shortening cycle of the DLMs¹⁶, and both increase as the DLMs are extended and decrease as they shorten. These spacing changes are in phase with each other and with the length cycle of the DLMs, suggesting that strains in the thick filaments are elastically coupled to sarcomere length, which is consistent with previous observations on passively stretched vertebrate muscle¹⁷. The amplitude of these cyclic strains, measured by comparing the 0.3 and 0.93 phase points (as indicated by two dashed lines in Fig. 3a), are different for the two reflections. The 14.5-nm reflection shows a 0.4% change (0.06 nm per repeat), twice the 0.2% change (0.014 nm per repeat) measured for the 7.2-nm reflection ($P < 0.05$). The larger movement of the 14.5-nm reflection suggests that the centres of mass, presumably the catalytic domains of the myosin heads, separate by an additional ~ 0.03 nm from each other as they are pulled towards the Z-disk by the thin filaments. Because the thin filaments are compliant, this shear force should be greatest near the Z-disk and diminish towards the M-line. This gradient would result in an increased separation of attached heads, which might explain the greater spacing change of the 14.5-nm reflection relative to the 7.2-nm reflection. However, this interpretation is complicated by the fact that the number of attached heads increases during extension, as discussed below. Nevertheless, the elastic deformation of the thick filaments during muscle lengthening immediately suggests a mechanism by which appreciable potential energy is stored in the muscle as elastic strain, available for conversion into kinetic energy during the shortening phase. Elastic storage has been proposed as an important feature of the flight motor that minimizes the inertial power costs of accelerating the wings back and forth in each wingbeat^{18–20}. These results would suggest that at least part of this storage takes place within the myofilaments themselves.

Unlike in *Lethocerus*, the 14.5-nm meridional reflection in *Drosophila* is weak in resting intact muscle, which is consistent with its relatively disordered appearance in electron micrographs²¹.

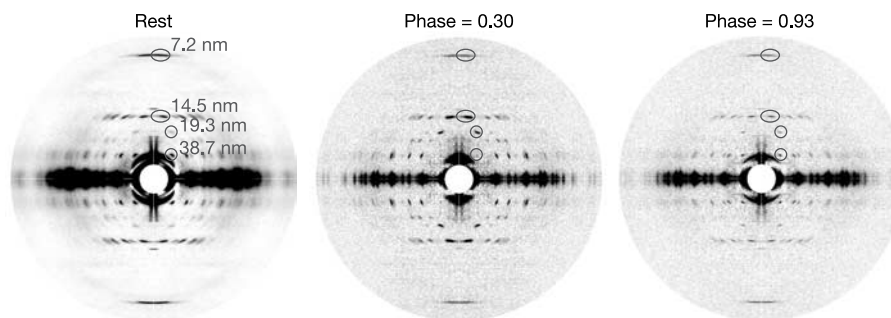


Figure 2 X-ray patterns from live flies. Shown are background-subtracted X-ray diffraction patterns from live *Drosophila metleri* at rest, near the end of lengthening (0.3 phase) and near the end of shortening (0.93 phase). Note the evident increase in overall ordering of the 0.3 phase with respect to the 0.93 phase (vertical dotted lines). The

positions of the 7.2-nm and 14.5-nm near-meridional reflections and the 38.7-nm and 19.3-nm reflections on the first row line are indicated. The 14.5-nm and 7.2-nm reflections in *Drosophila* sp. appear split because of axial stagger of thick filaments¹⁰. The resting pattern had a tenfold greater exposure.

Throughout the wingbeat cycle, the intensity of the 14.5-nm reflection varies by 400%, in phase with thick-filament strain (Fig. 3b). In contrast, the intensity of the 7.2-nm reflection undergoes only a 40% increase in intensity over the same period, indicating that the large change in the 14.5-nm reflection is not an artefact of the general ordering or alignment of the tissue. The intensity of the 14.5-nm reflection may be attributed to changes in the angle between the crossbridges and the long axes of the thick filaments, as well as to changes in the angular dispersion of the crossbridges with respect to that axis^{11,13,22}. The fourfold rise in the intensity of the 14.5-nm reflection can therefore be explained by the crossbridges becoming progressively more ordered and more

perpendicular to the thick-filament backbone during muscle lengthening. The corresponding decrease in intensity at 14.5 nm during shortening is consistent with the expected lever-arm tilting during crossbridge powerstrokes.

Figure 3c shows the variation in intensity of the 38.7-nm and 19.3-nm reflections on the first row line. The 38.7-nm reflection is thought to be due to the troponin complexes located every 38.7 nm along the thin filament with contributions from actin–myosin interactions in active muscle. In contrast, the 19.3-nm reflection is interpreted specifically as a marker for the interaction of myosin heads with the actin helix at sterically optimal target zones located midway between troponin repeats⁵. For stretch-activated *Lethocerus* IFM *in vitro*, it has been shown⁵ that myosin labelling at target zones increases the 19.3-nm reflection and diminishes the 38.7-nm reflection through destructive interference. Consistent with this is the observation that the 19.3-nm reflection in intact flies rises during stretching of the DLMs and declines during shortening, whereas the 38.7-nm reflection shows the opposite behaviour (Fig. 3c). The increase in intensity of the 19.3-nm reflection during lengthening provides direct evidence that a substantial fraction of the total number of cycling crossbridges form attachments at actin target zones as the DLMs are being stretched by the antagonistic dorsal-ventral muscles (DVMs). The importance of cycling crossbridges during lengthening in the underlying mechanics of stretch activation was first suggested by Squire²³ and accords with recent work²⁴ showing that some Ca^{2+} activated bridges must be attached to the thin filament before stretch activation can occur.

In both *Lethocerus* and *Drosophila* IFM, two isoforms of troponin C exist, one regulated by stretch (F1), the other by Ca^{2+} (F2)²⁵. Both are assumed to relieve the inhibitory effect of troponin I on cross-bridge formation by altering the position of tropomyosin to uncover actin-binding sites. The myosin heads that attach to actin during lengthening will be pulled towards the Z-disk, opposite to their direction of motion during powerstrokes. These strained crossbridges are therefore likely candidates for the 'sensors' that activate the stretch-dependent F1 isoform of troponin C, leading to the rapid binding of myosin heads at target zones as indicated by the observed changes in the 19.3-nm reflection intensity²⁴. Future time-resolved experiments that record the second actin layer line, a marker for tropomyosin movement during thin-filament activation, could serve to confirm this hypothesis.

During the transition from lengthening to shortening, muscle contraction must transiently approach isometric conditions (Fig. 3a). At this point, the 19.3-nm and 14.5-nm reflection intensities are maximal, indicating that a large proportion of heads are attached, well ordered and nearly perpendicular to the long axis of the thick filaments, an arrangement similar to isometrically contracting *Lethocerus* IFM *in vitro*²⁶ and to isolated frog muscle^{13,27}. The rapid decline in intensity at 14.5 nm at the onset of shortening suggests that the first crossbridge powerstrokes are roughly synchronous. In five of six preparations we detected a rapid decrease in the intensity of the 19.3-nm immediately after the onset of shortening, followed by a transient recovery (Fig. 3c). An interpretation of the initial decrease is that the first set of synchronously acting crossbridges detaches at the end of their powerstroke, followed by the attachment of another population, which immediately undergo powerstrokes that continue to shorten the muscle. Multiple powerstrokes are required during the shortening phase of the wingbeat because the relative axial displacement of the thick and thin filaments during shortening, about 3.5% or about 58 nm (ref. 16), is much longer than any single crossbridge step displacement (about 7 nm)²⁸. During shortening, the occupancy of target zones by bound crossbridges, as indicated by the intensity of the 19.3-nm reflection, remains high until roughly halfway through the downstroke, whereupon it decreases rapidly, concurrent with both the observed decrease in thick-filament strain (Fig. 3a) and the postulated decrease in strain of actin-bound crossbridges

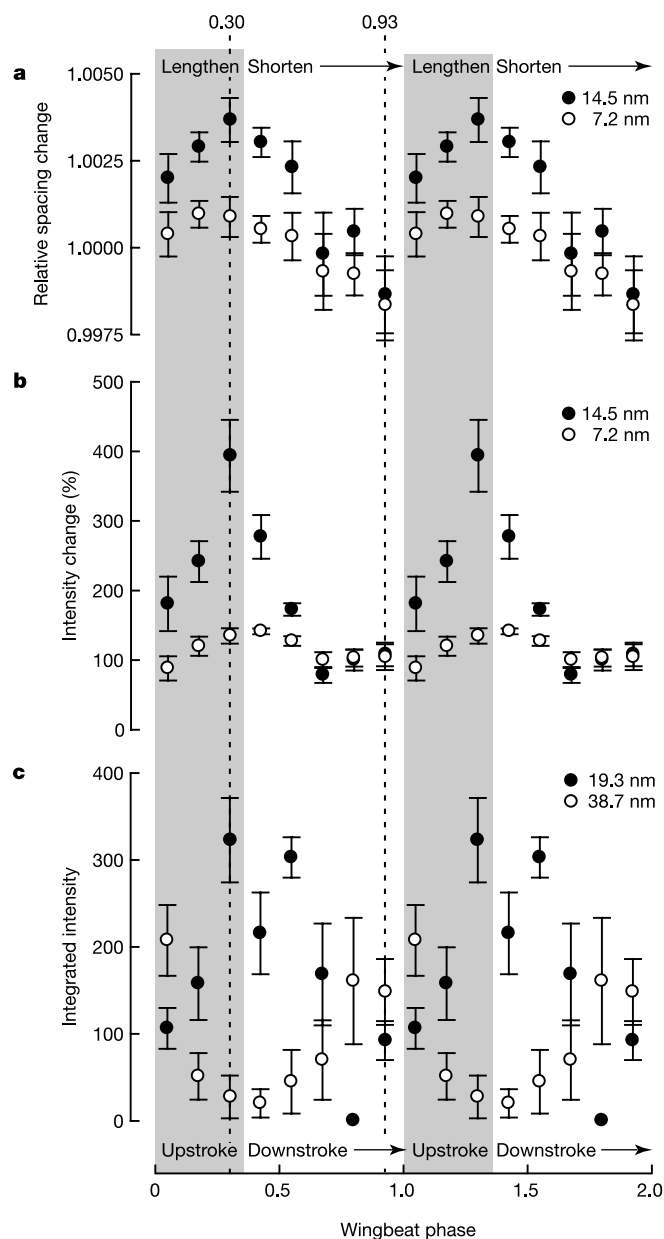


Figure 3 Changes in diffraction features as a function of phase of the wingbeat cycle. **a**, The 7.2-nm and 14.5-nm reflection spacings relative to rest. Thick-filament strain is approximately in phase with the wingbeat. **b**, The 7.2-nm and 14.5-nm reflection intensities relative to rest. A fourfold variation in the 14.5-nm reflection intensity is observed in phase with thick-filament strain. **c**, Integrated intensities of the first row line reflections on the 19.3-nm and 38.7-nm layer lines. Note the roughly reciprocal relation between the two intensities. Error bars show standard errors of the mean.

responsible for promoting stretch activation. In this view, it is the release in strain that causes all heads to detach abruptly, leading to shortening deactivation.

In these experiments we have coupled synchrotron X-ray diffraction with a tethered flight simulator to obtain an integrated picture of muscle structure and function within a living animal. The results of this time-resolved analysis shed light on two important features of the asynchronous flight muscle of insects. First, the strains measured within the stiff thick filaments might indicate a method by which lengthening flight muscles may store elastic energy for use in the next wingbeat. Second, changes in the number and orientation of crossbridge attachments throughout the wingbeat are likely to be instrumental in the process of stretch-activation. The crossbridge dynamics observed here provide constraints for future attempts to link the roughly 7-nm single myosin powerstrokes to the 50–60-nm movements of half sarcomeres within each 5-ms wingbeat cycle. Modifying the methods described here to enhance temporal resolution might help to address this question. *Drosophila* is ideally suited for these experiments, given its high degree of structural order and its suitability for future structure–function studies with genetic engineering methods²⁹. □

Methods

X-ray instrumentation

X-ray experiments used the small angle instrument on the BioCAT undulator-based beamline 18-ID (ref. 30) at the Advanced Photon Source, Argonne National Laboratory, Argonne, Illinois. The overall experimental arrangement is shown in Fig. 1. X-ray optics focus the beam horizontally at the fly and vertically at the detector. This optical arrangement allows the selection of these muscles to the exclusion of other nearby muscle systems. The X-ray beam energy was 12 keV (wavelength 0.103 nm), with a specimen-to-detector distance of 1.9 m. Diffraction patterns were recorded with a MAR 165 CCD detector (MARUSA). A rapid shutter set duration of exposure and beam intensity was adjusted appropriately in the range 10^{11} – 10^{13} photons s^{-1} with aluminium attenuators to obtain adequate counting statistics in the X-ray pattern with minimal radiation damage to the specimen.

Preparation of specimens

Flies (*Drosophila metleri* or *Drosophila melanogaster*) were grown under uncrowded conditions to increase the thickness of the muscle fibres intercepted by the X-ray beam. Before being mounted, flies (3–5-day-old females) were anaesthetized by being cooled to 4 °C with a thermoelectrically cooled stage. Flies were tethered by gluing a fine tungsten wire to the upper region of the thorax, permitting free movement of the head and wings during flight. Flies so tethered will spontaneously flap their wings for about 20–40 min. After recovery from anaesthesia (1 h), a fly was mounted with the long axis of its thorax perpendicular to the X-ray beam (head upwards) such that the 0.2-mm-wide collimated X-ray beam intersected the two parallel sets of DLM fibres located on both sides of the median plane of the thorax (Fig. 1). It was important to position the beam relatively low in the thorax so as to miss the fused thoracic–abdominal ganglion that contains the neuronal circuits driving the flight muscles, to reduce radiation damage to the nervous system and to reduce X-ray background from other structures.

Timing and synchronization

To maintain a steady wingbeat we used a ‘tethered flight simulator’ consisting of an optical wingbeat analyser that tracked the movements of the animal’s wings, and a panoramic electrical display that presented the animal with dynamic visual stimuli². Figure 1 shows a schematic diagram of the flight arena as installed in the X-ray diffraction apparatus. Using this apparatus, animals will fly for 30–120 min before they run out of energy stores. When placed in the arena, an infrared diode mounted at the top of the arena casts shadows of the animal’s wings on the two optical sensors of the wingbeat analyser, which track the back and forth motion of the wings within the wingbeat plane. From the raw analogue signals of these optical sensors, dedicated online circuitry determines the amplitude of both wings in each wingbeat as well as the instantaneous frequency. The visual display surrounding the animal consists of a cylindrical array of 16 light-emitting diode panels, each consisting of an 8×8 array of individual elements. Programmed images (bitmaps) are scanned onto the display at 1 kHz. In closed-loop mode, the behaviour of the animal (measured by the wingbeat analyser) is used to control the horizontal motion of the visual display that causes the fly to adjust its wingbeat frequency accordingly. The wingbeat analyser also provides a trigger pulse corresponding to a precise point in the wingbeat cycle about 1 ms before the start of the upstroke. A program running on a laptop computer was used to monitor the trigger pulse, to calculate the average wingbeat frequency from the running average of ten wingbeat periods and to trigger the shutter (opening time 340 μ s), after a delay set to a preset fraction (‘phase’) of the wingbeat period, once for every ten wingbeats. Diffraction patterns consisted of the sum of 150 such exposures (52.5 ms total exposure). There were eight phases per wingbeat, so each phase was separated by about 0.6 ms.

Data analysis

Intensities of meridional reflections were estimated by integrating the area under the peaks in one-dimensional projections with the multiple one-dimensional fitting routines in the program FIT2D (<http://www.esrf.fr/computing/scientific/FIT2D>) assuming a polynomial background and gaussian peaks. For estimation of the 38.7-nm and 19.3-nm first row line reflection intensities, a radially symmetric background was subtracted from each pattern by using the routines in the XFIX program in the CCP13 suite (<http://www.ccp13.ac.uk>) before analysis.

Received 20 August; accepted 29 November 2004; doi:10.1038/nature03230.

1. Josephson, R., Malamud, J. & Stokes, D. Asynchronous muscle: a primer. *J. Exp. Biol.* **203**, 2713–2722 (2000).
2. Lehmann, F. O. & Dickinson, M. H. The changes in power requirements and muscle efficiency during elevated force production in the fruit fly *Drosophila melanogaster*. *J. Exp. Biol.* **200**, 1133–1143 (1997).
3. Reedy, M. K., Holmes, K. C. & Tregear, R. T. Induced changes in orientation of the cross-bridges of glycerinated insect flight muscle. *Nature* **207**, 1276–1280 (1965).
4. Miller, A. & Tregear, R. T. Evidence concerning crossbridge attachment during muscle contraction. *Nature* **226**, 1060–1061 (1970).
5. Tregear, R. T. *et al.* X-ray diffraction indicates that active cross-bridges bind to actin target zones in insect flight muscle. *Biophys. J.* **74**, 1439–1451 (1998).
6. al-Khayat, H. A., Yagi, N. & Squire, J. M. Structural changes in actin-tropomyosin during muscle regulation: computer modelling of low-angle X-ray diffraction data. *J. Mol. Biol.* **252**, 611–632 (1995).
7. Reedy, M. K. & Reedy, M. C. Rigor crossbridge structure in tilted single filament layers and flared-X formations from insect flight muscle. *J. Mol. Biol.* **185**, 145–176 (1985).
8. Taylor, K. A. *et al.* Tomographic 3D reconstruction of quick-frozen, Ca^{2+} -activated contracting insect flight muscle. *Cell* **99**, 421–431 (1999).
9. Reedy, M. C. Visualizing myosin’s power stroke in muscle contraction. *J. Cell Sci.* **113**, 3551–3562 (2000).
10. Irving, T. C. & Maughan, D. W. *In vivo* x-ray diffraction of indirect flight muscle from *Drosophila melanogaster*. *Biophys. J.* **78**, 2511–2515 (2000).
11. Huxley, H. E. *et al.* Changes in the X-ray reflections from contracting muscle during rapid mechanical transients and their structural implications. *J. Mol. Biol.* **169**, 469–506 (1983).
12. Irving, M., Lombardi, V., Piazzesi, G. & Ferenczi, M. A. Myosin head movements are synchronous with the elementary force-generating process in muscle. *Nature* **357**, 156–158 (1992).
13. Irving, M. *et al.* Conformation of the myosin motor during force generation in skeletal muscle. *Nature Struct. Biol.* **7**, 482–485 (2000).
14. Huxley, H. E., Stewart, A. & Irving, T. Spacing changes in the actin and myosin filaments during activation, and their implications. *Adv. Exp. Med. Biol.* **453**, 281–288 (1998).
15. Huxley, H. E., Reconditi, M., Stewart, A. & Irving, T. What the higher order meridional reflections tell us. *Biophys. J.* **84**, 139a (2003).
16. Chan, W. P. & Dickinson, M. H. *In vivo* length oscillations of indirect flight muscles in the fruit fly *Drosophila virilis*. *J. Exp. Biol.* **199**, 2767–2774 (1996).
17. Wakabayashi, K. *et al.* X-ray diffraction evidence for the extensibility of actin and myosin filaments during muscle contraction. *Biophys. J.* **67**, 2422–2435 (1994).
18. Weis-Fogh, T. in *XV Int. Congr. Entomol. (Royal Entomological Society, London, 8 to 16 July 1964)* (ed. Freeman, P.) 186–188 (XII International Congress of Entomology, London, 1965).
19. Alexander, R. M. & Bennet-Clark, H. C. Storage of elastic strain energy in muscle and other tissues. *Nature* **265**, 380–390 (1977).
20. Dickinson, M. H. & Lighton, J. R. Muscle efficiency and elastic storage in the flight motor of *Drosophila*. *Science* **268**, 87–90 (1995).
21. Menetret, J. F., Schroder, R. R. & Hofmann, W. Cryo-electron microscopic studies of relaxed striated muscle thick filaments. *J. Muscle Res. Cell Motil.* **11**, 1–11 (1990).
22. Dobbie, I. *et al.* Elastic bending and active tilting of myosin heads during muscle contraction. *Nature* **396**, 383–387 (1998).
23. Squire, J. M. *The Structural Basis of Muscular Contraction* (Plenum, New York, 1981).
24. Linari, M., Reedy, M. K., Reedy, M. C., Lombardi, V. & Piazzesi, G. Ca-activation and stretch-activation in insect flight muscle. *Biophys. J.* **87**, 1101–1111 (2004).
25. Agianian, B. *et al.* A troponin switch that regulates muscle contraction by stretch instead of calcium. *EMBO J.* **23**, 772–779 (2004).
26. Tregear, R. T. *et al.* Cross-bridge number, position, and angle in target zones of cryofixed isometrically active insect flight muscle. *Biophys. J.* **86**, 3009–3019 (2004).
27. Juanhuix, J. *et al.* Axial disposition of myosin heads in isometrically contracting muscles. *Biophys. J.* **80**, 1429–1441 (2001).
28. Littlefield, K. P. *et al.* The converter domain modulates kinetic properties of *Drosophila* myosin. *Am. J. Physiol. Cell Physiol.* **284**, C1031–C1038 (2003).
29. Sparrow, J., Drummond, D., Peckham, M., Hennessey, E. & White, D. Protein engineering and the study of muscle contraction in *Drosophila* flight muscles. *J. Cell Sci. Suppl.* **14**, 73–78 (1991).
30. Irving, T. C., Fischetti, R. E., Rosenbaum, G. & Bunker, G. B. Fiber diffraction using the BioCAT Undulator Beamline at the Advanced Photon Source. *Nucl. Instrum. Methods A* **448**, 250–254 (2000).

Supplementary Information accompanies the paper on www.nature.com/nature.

Acknowledgements We thank J. Costello for help with data analysis, J. Fockler for computer programming, and D. Swank, R. Tregear, M. K. Reedy and M. C. Reedy for helpful discussions. The research was supported by NIH. The APS is supported by the US Department of Energy. BioCAT is a NIH-supported Research Center.

Competing interests statement The authors declare that they have no competing financial interests.

Correspondence and requests for materials should be addressed to T.I. (irving@iit.edu).

Contacts

Publisher: Ben Crowe
Editor: Paul Smaglik
Marketing Manager: David Bowen

US Head Office, New York

345 Park Avenue South, 10th Floor,
New York, NY 10010-1707
Tel +1 800 989 7718
Fax +1 800 989 7103
e-mail: naturejobs@natureny.com

US Sales Manager/ Corporations:

Peter Bless

Classified Sales Representatives:

Phone: +1 800 989 7718

New York/ Pennsylvania/

Latin America:

Kelly Roman

Midwest USA/ Maryland/ NIH:

Wade Tucker

East USA/ Canada:

Janine Taormina

San Francisco Office

Classified Sales Representative:

Michaela Bjorkman

West USA/ West Corp. Canada

225 Bush Street, Suite 1453

San Francisco, CA 94104

Tel +1 415 781 3803

Fax +1 415 781 3805

e-mail: m.bjorkman@nature.sf.com

European Head Office, London

The Macmillan Building, 4 Crinan Street,
London N1 9XW, UK

Tel +44 (0) 20 7843 4961

Fax +44 (0) 20 7843 4996

e-mail: naturejobs@nature.com

Naturejobs Sales Director:

Nevin Bayoumi (4978)

Advertising Production

Manager: Billie Franklin

To send materials use London
address above.

Tel +44 (0) 20 7843 4814

Fax +44 (0) 20 7843 4996

e-mail: naturejobs@nature.com

Naturejobs web development:

Tom Hancock

Naturejobs online production:

Niamh Shields

European Satellite Office

Germany/ Austria/ Italy/

The Netherlands/ Belgium:

Patrick Phelan

e-mail: p.phelan@nature.com

Sharon de Weert

e-mail: s.deweert@nature.com

Japan Head Office, Tokyo

MG Ichigaya Building (5F),

19-1 Harakatamachi,

Shinjuku-ku,

Tokyo 162-0841

Tel +81 3 3267 8751

Fax +81 3 3267 8746

Asia-Pacific Sales Director:

Rinoko Asami

e-mail: rasami@naturejpn.com

naturejobs

Closing the gap

In the life sciences, the problems of graduate training are legion, and well-documented. PhDs take an increasingly long time to complete, graduate programmes often don't account for true interdisciplinary learning, and the gap between basic and applied science remains wide. But some key institutions seem to be paying attention: over the past few months, a number of programmes and initiatives have been announced that aim to address these issues.

The Howard Hughes Medical Institute, for example, has launched two initiatives to fund training programmes — one for interdisciplinary science and another for integrating medicine into biomedical training. The US National Institutes of Health will fund the studentships for institutions that receive these grants.

The Memorial Sloan-Kettering Cancer Center in New York is taking a similar approach for its new graduate school, announced last November. This school will train young scientists in basic research, but with an eye for applications in cancer. Getting students to interact with both basic scientists and clinicians will improve the kinds of questions they ask, and the research approaches they take, says Sloan-Kettering's president, Harold Varmus.

And the University of Pennsylvania last week announced it was setting up three institutes that would take a similar approach to integrating biomedical research, education and patient care.

Many of these programmes have taken the lead from pioneers in interdisciplinary education, such as the European Molecular Biology Laboratory, by having students pass through multiple labs and ensuring that graduate students have good mentoring. Ideally, the new programmes adopting these approaches should become commonplace, supplanting earlier complaints about graduate education with more and more success stories.

Paul Smaglik
Naturejobs editor



Contents

POSTDOCS AND STUDENTS

Planning for the future p336

CAREER VIEW

Scientists & Societies

International Institute for

Applied Systems Analysis

Graduate Journal

Time to explore new worlds

Movers

Iain Mattaj

p338

WWW.NATUREJOBS.COM

Career centre

Information on the
scientific job market

FOCUS

SPOTLIGHT

RECRUITMENT

ANNOUNCEMENTS

EVENTS

Save now, don't pay later

Should young scientists be tightening their belts to save for the future? Kendall Powell compounds the interest.

As a 35-year-old junior faculty member at the Mayo Clinic in Rochester, Minnesota, David Katzmman is relieved to be finally starting his retirement savings in earnest. While he was a doctoral student and postdoc, Katzmman had no real savings and a house was his only investment.

And he is far from alone. Most young scientists admit to a state of financial denial between leaving university and landing their first permanent position. That, on average, takes 7–10 years according to the US National Science Foundation's Survey of Doctorate Recipients. And waiting to save later on a higher salary is a big gamble: it assumes that no financial crisis will occur and that training will not stretch beyond a decade.

"Frankly, most scientists don't think about this stuff because it's not on the radar," Katzmman says. "I would be happy to never have to think about money." Katzmman plans to catch up on his retirement savings now by tripling his effort. But he knows that waiting to save until later, like most people do, has meant forfeiting precious years of compounding interest.

Some money-minded young researchers have come up with savings strategies even on modest pay cheques. Most are strikingly simple, but they require discipline and all-important consistency. And as the experts tell us, everyone should be saving something — even if it seems like small potatoes.

FUTURE RETURNS

Young researchers who tend to make what other professions would call 'starting pay' for almost a decade — ranging from \$14,000 graduate stipends to \$40,000 for experienced postdocs — usually feel that saving money requires difficult choices between today's lifestyle and tomorrow's security. Many choose to focus on more manageable short-term goals, such as buying a car. Or they feel capable of working towards just one long-term goal and choose between buying a house, children's education funds or retirement.

And many scientists say that they just don't have the knowledge, time or mental energy that financial planning requires. Faced with these challenges, many adopt a 'why bother' attitude, choosing to concentrate on career development to secure their future.

As a graduate student, David Goetz says he followed the "stingy strategy" of living really cheaply to save about \$25,000. But in 2001, his last year of graduate school, he watched it disappear in the stock-market crash. "It wasn't a large enough amount of money to diversify my investments," he says.

So when he started a postdoc at the University of California, San Francisco, Goetz says that he originally had no investment plan. But his cheap living led to



money piling up in his account by default. Three years later, he says, he has to decide what to do with the extra money: "Buy a new car? Keep it in a savings account with low yield? Or something different. That's the question I'm debating now."

Because California real estate is out of the question, Goetz is thinking of buying a house in San Antonio, Texas — an investment less volatile than the stock market and which at least gives him a tax break. He knows other postdocs who are also considering buying out-of-town houses, but admits it is still financially a bit risky. "As a postdoc you don't have much cushion if something happens," such as major damage from tenants, he says.

Heather Leslie and her husband Jeremy Rich, both postdocs at Princeton University in New Jersey, have taken the simple-but-disciplined approach to their retirement savings. They have continued to live like graduate students in subsidized university housing with donated furniture. That way, "a large fraction of our joint monthly income goes to savings", she says. They also revisit their individual retirement accounts (IRAs) if any money is left over after taxes.

"My father always stressed letting your money grow through time," Leslie says. The couple, like many

young scientists, started their nest eggs with money from a family inheritance. Leslie recommends checking out large investment companies that have low fees, low buy-in thresholds and easy web-based customer service for busy lab workers. Adam Mullick, president of the Society of Fellows of the Scripps Research Institute at La Jolla, California, couldn't agree more: he recommends saving at least the allowed, tax-deductible \$4,000 maximum per year in an IRA.

"When you look at how much growth depends on when you start contributing, it's really foolish not to start as early as possible," says Mullick. Leslie says her next goal is to have three months of income set aside for a safety-net fund. This can be especially important because postdoctoral funding can be unpredictable or short-term.

THE EXCHANGE RATE

Short-term contracts and a more frequent change of job, especially moving between different countries, give European scientists-in-training a somewhat different set of financial problems. As foreign workers in 'temporary' positions, many cannot take full advantage of the usual social benefits, such as unemployment and pension schemes, and have a difficult time getting loans to buy property.

"We consider this stretch of ten years between graduation and a position to be very unpleasant for young researchers in Europe in general," says Renzo Rubele, a doctoral candidate at the University of Salerno in Italy and president of the EURODOC professional organization for graduate students and postdocs. He says it is extremely rare for European researchers to save money beyond their monthly expenses. But, he says, "the culture is that you don't complain if you think that in the near future you will have a permanent position."

Agnė Kulytė, a 31-year-old Lithuanian postdoc working at the Karolinska Institute in Stockholm, Sweden, says that although she has saved a bit of an emergency fund, she feels that having children and buying an apartment are financially out of reach. Although this is a typical situation, she is frustrated by the nagging financial worries. "You cannot live your life thinking 'I'm not going to get sick, or be unemployed, or I'll worry about retirement later'," she says.

"But what really hurts is that my education just doesn't correspond to the income," Kulytė adds. And a more stable, higher-paying job may not come up for many European scientists until middle age, when it may be too late to have children or catch up on savings. Another Karolinska postdoc, Therese Pham, who is 37, says that she definitely feels she is falling behind her peers in other professions.

And Annika Armulik, also at the Karolinska, says she has moved from Estonia to Germany to Sweden for graduate school and postdoc work, and will soon move to Switzerland in search of another position. She worries that the different national pension funds set

up for her may not transfer to her when she retires in perhaps yet another country. "I am 34 now and feel I have nothing behind me," she says.

ATTITUDE ADJUSTMENT

Financial advisers and money-management experts argue that there's no reason for any young scientist to feel that way. They say it is possible to invest wisely whether you are a graduate student making \$15,000 a year or a tenured investigator making \$150,000. And delaying financial planning and saving until after the postdoc stage makes little sense.

"That is a mindset that the scientific community needs to break," says Dave Ramsey, a money-

management talk-radio host based in Nashville, Tennessee. He points to the fact that investing \$100 per month from age 20 to 60 results in just over a million dollars, if you assume historic rates of return. But the same amount saved from age 35 to 60 will give you only about \$170,000. And even if the amount you save is tripled to \$300 from age 35 to 60, it still won't catch you up, giving you only about \$500,000. "It's the cost of pizza and cable to retire with some dignity rather than retiring broke," he says.

Ramsey also recommends that young scientists try to live within, or even below, their means. In other words, a graduate student should drive a paid-for \$4,000 car rather than making payments on a \$17,000 one. He says everyone should prioritize their savings efforts into attainable goals, so that as one goal is met, the next is clear. A high priority would be to establish and maintain an emergency fund of 3–6 months of income. After that, 15% of income should go towards retirement savings.

And after that, concentrate on long-term, large investments, such as a mortgage or children's education funds. In a bit of good news for young scientists trying to juggle expenses, Ramsey says that it's not necessary to save up a large sum of money before you have children. Sure, they add to monthly expenses, he says, but it's not substantial enough to require saving in advance.

Most young scientists find themselves without the benefit of matched savings plans from their employers or even good financial guidance. But it's not impossible to find your way to a solid money-management plan through a consultation with a financial adviser, or through a university benefits office or credit union.

The Scripps Research Institute offers all employees lunchtime seminars on everything from demystifying your credit rating to buying a house to how to invest.

Nevertheless, many young researchers will continue to play and work hard now, leaving savings to later. And if so, there's advice for that group as well. When Katzmann began his investigator position at Mayo, he set up the automatic pay deductions for his pension plan right away. "It's so much easier to never see that money. In going from a postdoc salary to a principal-investigator salary, missing \$300 a month is nothing. Set it up before you get used to that money."

Kendal Powell is a freelance science writer based in Broomfield, Colorado.



Renzo Rubele worries that younger researchers don't save enough money.



Well prepared: David Katzmann is keen to get saving for his retirement.

GRADUATE JOURNAL

Time to explore new worlds

As 2005 begins, it is gratifying but also disconcerting to say that this year, I will graduate. I'll complete a PhD in molecular biology at the University of California, Los Angeles. But despite this impending accomplishment, I can honestly say that at 30 years old I still don't know what I want to be when I grow up.

For the past eight years or so, I've studied aspects of mRNA in mammalian gene regulation. After taking some dead-end paths, I am quite happy with the stories that my dissertation research will tell. I began my graduate studies with the same fervour for RNA biology that I now possess. If I based my next career choice solely on following my scientific interest, I'd continue working with RNA and stagger down that well-worn path to postdoc and, if I'm lucky, professor.

But, life is not all about the lab and my love stretches beyond ribose chains. I'm a happily married man and my wife is a disturbingly brilliant public-policy researcher. Looking into how we'll solve our two-job dilemma, I'm exploring science writing and public policy. Should I test those waters or study a new problem at the bench? All have their appeal and I'll be learning about these paths as I finish my degree. This is going to be an interesting year. ■

Jason Underwood is a graduate student in microbiology at the University of California, Los Angeles.

SCIENTISTS & SOCIETIES

International Institute for Applied Systems Analysis, Austria

Hardly anyone could believe my summer plans. "You're going to work in an Austrian castle?" a fellow graduate student asked. "Well, it's more like a palace," I explained, somewhat sheepish about planning to research civil war in such luxurious surroundings.

The truth was I couldn't believe my luck. After completing my second year of graduate studies in demography at the University of California, Berkeley, I was accepted at the International Institute for Applied Systems Analysis (IIASA), a small, multilateral, autonomous research institution that runs a summer programme for PhD students. Located just outside Vienna, the IIASA allows scientists from all over the world to convene and collaborate to address issues within the

institute's three theme areas: natural resources and environment; population and society; and energy and technology.

Each summer, about 60 graduate students work on their PhDs through IIASA's Young Scientists Summer Programme (YSSP). This offers the students a chance to explore the policy implications of their work. The IIASA's projects range from the technical — developing and applying mathematical modelling techniques — to the practical — integrating social and natural-science models and data in policy-relevant analyses.

I spent most of my time on my research into whether a young population and abundance of natural resources make civil wars more likely to start. My mentor gave me helpful input on how to model these processes over time. Seminars and focus-group discussions exposed me to the work of important

worldwide scholars.

A midsummer workshop allowed us to share our current research. Many of my YSSP colleagues were environmental scientists, geographers, political scientists and biologists, so the interdisciplinary feedback I received was very helpful. Students and institute scholars suggested articles, techniques and data sets I might never have encountered otherwise.

The YSSP helped me look at my research from fresh angles. It offered the springboard I needed to move from coursework to productive independent research. I've returned with a potential dissertation chapter, lots of new ideas on how to proceed, and a new network of colleagues, friends and potential collaborators from around the world. ■

Sarah Elizabeth Staveteig is a graduate student at the University of California, Berkeley.

♦ www.iiasa.ac.at/Admin/YSPP/yssp2005/about-program.html

MOVERS **Iain Mattaj, director-general, European Molecular Biology Laboratory, Heidelberg, Germany**

Frustrated by the slow progress of his work at the beginning of his postgraduate career, Iain Mattaj once considered joining a friend's retail business. But his desire to understand how the world functions at the molecular level won, thanks in part to an influential mentor. In May, the 52-year-old molecular biologist will take the reins of the European Molecular Biology Laboratory (EMBL) in Heidelberg, Germany, one of Europe's premier research institutes.

Mattaj's understanding of science was influenced by the enthusiasm and creativity of developmental biologist Eddy de Robertis, his supervisor at the

Biocentre Basel in Switzerland: "He taught me to constantly think about the most important question one could ask."

In 1985, when de Robertis joined the University of California, Los Angeles, the Scotsman Mattaj took the opportunity to set up his own group at EMBL, where a new gene-expression programme had just been started. He never regretted the decision. His research in the late 1980s, one of his most productive periods, resulted in a solid understanding of how macromolecules are trafficked between the nucleus and the cytoplasm.

At the time, Mattaj recalls, things still moved slowly. "Nowadays, researchers live in a whirlwind of information," he says. He is well aware that the pace has increased pressure, in particular on young scientists in highly competitive fields. Graduates considering a scientific career should therefore think very well about the question they want to have answered, and then carefully chose the

right place to go, he says. "It is crucial to learn in a top-quality environment."

When he starts as EMBL director, Mattaj can rely on long experience in science management. After chairing the gene-expression programme for ten years, he was promoted to be EMBL's scientific director in 1999.

Looking back, he appreciates most that as a young scientist he could do independent research with very little red tape. This was only possible, he says, because senior colleagues were looking after the administration. Now he wants to do the same for the next generation. "I find satisfaction in organizing things for the community," he says, "even though it is rarely so much fun as doing science."

But he is keen to stay as closely in touch with research as possible. "We're just beginning to understand how the nucleus is built and organized," he says. There is a good chance that the almost-retailer will know the answer soon. ■

CV

1999–2005: Scientific director of the European Molecular Biology Laboratory (EMBL), Heidelberg, Germany.

1990–99: Programme coordinator, gene-expression programme, EMBL, Heidelberg, Germany.

1985–90: Group leader, EMBL, Heidelberg, Germany.